The 12th Annual International Conference on Research in Computational Molecular Biology (RECOMB2008)

30th March – 2nd April 2008

University Cultural Centre
National University of Singapore

Poster Book

Preface

This year, we received a number of quality poster abstracts with authors coming from a total of 28 different countries or regions---Australia, Belgium, Canada, China, Denmark, Estonia, France, Germany, India, Italy, Iran, Israel, Japan, Korea, Mexico, Netherlands, Poland, Russia, Singapore, South Africa, Sweden, Switzerland, Taiwan, Thailand, Turkey, UAE, UK, and USA. The abstracts were briefly reviewed by the poster committee members. In total 146 posters were accepted. We would like to thank all the authors who submitted posters and participated in RECOMB2008.

Poster Committee

Hon-Nian Chua, *Institute for Infocomm Research, Singapore*
Neil Clarke, *Genome Institute of Singapore, Singapore*
Frank Eisenhaber, *Bioinformatics Institute, Singapore*
Wing-Kin Sung, *National University of Singapore, Singapore*
Greg Tucker-Kellogg, *Lilly Singapore Centre for Drug Discovery, Singapore*
Martin Vingron, *Max Planck Institute for Molecular Genetics, Germany (Co-chair)*
Limsoon Wong, *National University of Singapore, Singapore (Co-chair)*

Accepted Posters

P2/ V. S. Gomase, S. V. Gomase, S. Tagore, D. A. Bhiwgade, M. M. V. Baig, K. V. Kale. *Computer-Aided Multi-Parameter Antigen Design: Impact of Synthetic Peptide Vaccines*

P3/ V. S. Gomase, S. V. Gomase, S. Tagore, D. A. Bhiwgade, M. M. V. Baig, K. V. Kale. *Prediction of MHC Binder for Fragment-Based Viral Peptide Vaccines from Cabbage Leaf Curl Virus*

P4/ Madhu Bala Priyadarshi. *Plant DNA Fingerprint Software*

P5/ David Wood, Mhairi Marshall, Shuzhi Cai, Matthew Bryant, David Hansen, Dominique Gorse. *Platform for Integrated and Accessible Bioinformatics*

P6/ Susan Tang, Fiona C. L. Hyland, Tomas C. Wessel, Jon Sorenson, Heather Peckham, Francisco M. De La Vega. *DiBayes: A SNP Detection Algorithm for Next-Generation Dibase Sequencing*

P7/ Yu Xue, Jian Ren, Longping Wen, Xuebiao Yao. *GPS 2.0: Prediction of Kinase-Specific Phosphorylation Sites in Hierarchy*

P8/ Jaya Iyer, Arathi Raghunath, Jignesh Bhate. *Analyzing the Role of CREBBP Co-Activator in Erythroid Differentiation—a Literature Mining Approach*

P9/ Monika Koul, Megha Kadam, Yashpal Malik, Ashok Kumar Tiwari, Jawaharlal Vegad. *Nucleocapsid Gene Sequence Analysis and Characterization of an Indian Isolate of Avian Infectious Bronchitis Virus*

P10/ Lesheng Kong, Alan Christoffels. *Large-Scale Gene Duplication Detection in Ciona savignyi and Ciona intestinalis*

P11/ Thanneer Malai Perumal, Wu Yan, Rudiyanto Gunawan. *In Silico Dynamical Analysis of Cellular Systems: A Molecular Perturbation Approach*

P12/ Marta Szachniuk, Marek Blazewicz, Mariusz Popenda, Ryszard W. Adamiak. *Design of RNA Fragments Structural Database*

P13/ Hongseok Tae, Kiejung Park. *ASMPKS: An Analysis System for Modular Polyketide Synthases*

P14/ Saurabh Shukla, Alok Shekhar. *In-silico Analysis of HIV-1 P1 Sequence and Structure Prediction*

P15/ Paul D. Yoo, Yung Shwen Ho, Bing Bing Zhou, Albert Y. Zomaya. *SiteSeek: Phosphorylation Site Predictor Using Adaptive Locality-Effective Kernel Methods and New Sequence Profiles*

P16/ Nandita Das, Vaibhav Navaghare, Vidyendra Sadanandan, Jignesh Bhate, Jaya Iyer. *Exploring Protein-Protein Interactions at the Domain Level*

P17/ Sangjukta Kashyap, Nandita Das, Usha Mahadevan, Jignesh Bahate. *In-silico Disease Target Screening and Evaluation with RNAi Data*

P18/ Victor Tomilov, Valery. Chernukhin, Murat Abdurashitov, Danila Gonchar, Sergei Degtyarev. *Cleavage of Mammalian Chromosomal DNA by Restriction Enzymes In Silico*

P20/ Murat Abdurashitov, Victor Tomilov, Valery Chernukhin, Danila Gonchar,Sergei Degtyarev. *Comparative Analysis of Human Chromosomal DNA Digestion with Restriction Endonucleases In Vitro and In Silico*

P21/ Sebastian Maurer-Stroh, Petra Van Damme, Joost Van Durme, Kim Plasman, Evy Timmerman, Pieter-Jan De Bock, Marc Goethals, Frederic Rousseau, Joost

Schymkowitz, Joel Vandekerckhove, Kris Gevaert. *Granzyme B Cleavage Site Predictions based on Sequence, Physical Property and Structural Description of the Motif*

P22/ Javed Mohammed Khan, Shoba Ranganathan. *A Multi-Species Comparative Structural Bioinformatic Analysis of Inherited Mutations in α-D-Mannosidase*

P23/ Diane Simarmata, Joo Chuan Tong, Philippe Kourilsky, Lisa F.P. Ng. *Large-Scale Analysis and Screening of Chikungunya Virus T-cell Epitopes*

P24/ Shunsuke Kamijo, Akihiko Fujii, Kenji Onodera, Kenichi Wakabayshi, Takatsugu Kobayashi, Kensaku Sakamoto. *Statistical Analysis of KMSKS Motif in Aminoacyl-tRNA Synthetase by Building a Library of Random Sequences*

P26/ Jacek Blazewicz, Marcin Bryja, Marek Figlerowicz, Piotr Gawron, Marta Kasprzak, Darren Platt, Jakub Przybytek, Aleksandra Swiercz, Lukasz Szajkowski. *Whole Genome Assembly from 454 Sequencing Output*

P27/ Hongseok Tae, Kiejung Park. *ConView: An Easy and Fast Visualization Tool for Contig Assembly*

P28/ Bhakti Bhagwat, Santosh Atanur, Sunitha Manjari, and Rajendra Joshi. *Deciphering Functional Linkages between Mycobacterium Tuberculosis H37Rv Proteins via Gene Ontology Similarity Scores*

P29/ Bruno Schwenk, Joachim Selbig, Matthias Holschneider. *Planes in 3-Dimensional Metabolite Triplet Data: A Robust, Bayesian Approach*

P30/ Tara Hessa, Nadja M. Meindl-Beinker, Andreas Bernsel, Hyun Kim, Yoko Sato, Mirjam Lerch-Bader, IngMarie Nilsson, Stephen H. White, Gunnar von Heijne. *Deciphering Transmembrane Helix Recognition by the ER Translocon*

P31/ Jens-Uwe Krause, Juergen Kleffe. *BACOLAP: BAC Assembly based on Bit-Vectors*

P32/ Rohit Reja, Venkata Krishnan AJ, Sandeep Kumar Yelakanti, Vishal Kumar Nangala, Umesh Roy, Rajib Bandopadhyay, Prashanth Suravajhala. *Host Pathogenesis and Lateral Gene Transfer Revisited: Challenges to Evolution*

P33/ Duangdao Wichadakul, Supawadee Ingsriswang, Eakasit Pacharawongsakda, Boonyarat Phadermrod, Sunai Yokwai. *ATGC-Dom: Alignment, Tree, and Graph for Comparative Proteomes by Domain Architecture*

P34/ Guimei Liu, Jinyan Li, Suryani Lukman, Limsoon Wong. *Predicting Protein Interactions Using Interacting Motif Pairs*

P36/ Jacek Blazewicz, Edmund K. Burke, Marta Kasprzak, Alexandr Kovalev, Mikhail Y. Kovalyov. *Dynamic Programming and Approximation Algorithms for the Simplified Partial Digest Problem*

P37/ Yew Chung Tang, Gregory Stephanopoulos, Heng-Phon Too. *Information-Theoretic Analysis for Exploring Cell Death-Survival Signaling*

P38/ Marta Szachniuk, Mariusz Popenda, Lukasz Popenda, Jacek Blazewicz. *Constructing Transfer Pathways in Multidimensional NMR Spectra of RNAs*

P39/ Bo Kim, Bruce Tidor, Jacob White. *Robust Optimization for Biological Network Calibration*

P40/ Pablo Carbonell, Antonio del Sol. *Specificity and Affinity of Protein-Protein Interactions: From a Systems Biology to a Molecular Point of View*

P41/ Dong Difeng, Limsoon Wong. *Using Gene Expression Analysis for Drug Pathway Identification: An Example on Nasopharyngeal Carcinoma (NPC)*

P42/ Geoffrey Koh, David Hsu, P.S Thiagarajan. Composition of Signaling Pathway Models and its Application to Parameter Estimation

P43/ Kazuo Iida, Takako Takai-Igarashi, Daiya Takai, Hiroshi Tanaka. *Knowledge-Based Pathway Optimization Strategy for Gene Expression Profiling Analysis*

P44/ Julia Medvedeva, Marina Fridman, Nina Oparina, Dmitri Malko, Ekaterina Ermankova, Ivan Kulakovsky, Vsevolod Makeev. *Reduced CpG Mutation Rate Suggests Functional Role of Intragenic and 3 CpG Islands in HumanGenes*

P45/ Saboura Ashkevarian, Armin Madadkar Sobhani, Bahram Goliaei. *Structural Modeling of MaSp1 and MaSp2 Proteins of Dragline Silk in Latrodectus hesperus*

P47/ Kyungsook Kim, Mira Oh, Jangsun Baek, Young Sook Son. *Missing Values Estimation for DNA Microarray Gene Expression Data: SPLS*

P48/ Leonard Lipovich, Charlie W-H. Lee, Hui Jia, Yuri Orlov, Thomas Wee-Hong Ng, Jieming Chen, Edwin Lian-Chong Ng, Edison T. Liu, Lance D. Miller, Lawrence W. Stanton, Ken W.-K. Sung, Vladimir A. Kuznetsov. *Analysis of Human Cis-Antisense Transcription: Primate-Specific Exonic Sequences, Structure-Dependent Sense-Antisense Co-Expression, and functionally Restricted Noncoding-RNA Transcription*

P49/ Rajeev Kumar, Manmath Routray, J. Febin Prabhu Dass. *HIV-1 Protease Inhibitor Comparative Docking Studies of Synthetic and Natural Compounds*

P50/ Andrea Sackmann, Piotr Formanowicz, Jacek Blazewicz. *A DNA-Based Algorithm for Calculating the Maxflow in Networks*

P51/ Mira Oh, Kyungsook Kim, Young Sook Son. *Neural Network Imputation for Missing Values in Time-Course Gene Expression Data*

P52/ Serban Nacu. *Gene Expression Network Analysis*

P53/ Chao Xie, Martti T. Tammi. *Discovery of DNA Copy Number Variation Using Shotgun Sequencing Data*

P54/ Juntao Li, Lei Zhu, Majid Eshaghi, Jianhua Liu, R. Krishna Murthy Karuturi. *Genome-Wide High-Density ChIP-Chip Tiling Array Data Analysis in Fission Yeast*

P55/ Wong Chee-Hong, Ooi Hong Sain, Georg Schneider. *The ANNOTATOR Software Environment: A Flexible Sequence Analysis Platform*

P56/ Sumantra Chatterjee, Guillaume Bourque, Thomas Lufkin. *A Bioinformatic and Transgenic Approach for Elucidating Tissue Specific Regulatory Elements*

P57/ Rosaura Palma-Orozco, Jorge Luis Rosas-Trigueros. *Cellular Automata and Simulation of Biological Processes*

P58/ Rory Johnson, Galih Kunarso, Christina Teh, Kee-Yew Wong, Kandhadayar G. Srinivasan, Sarah S.-L. Chan, R. Krishna Murthy Karuturi, Leonard Lipovich, Noel J. Buckley, Lawrence W. Stanton. *Genomic Analysis of Transcriptional Regulation by the Factor REST in Embryonic Stem Cells*

P59/ Mohammed A. Khidhir, K. Praveen Kumar, Marwa Al-Aseer. *Genetic Characterization and Population Structure of Arabian Tahr (Hemitragus jayakari) based on Microsatellites Analysis*

P60/ Menaka Rajapakse, Lin Feng. *Guided-Discovery of Motifs for Peptide Binding Prediction*

P61/ Xiao-Li Li, See-Kiong Ng. *Mining for Domain Dependency Sets from Protein Interactions*

P62/ Stephen Rumble, Michael Brudno. *SHRiMP: The Short Read Mapping Package*

P63/ Seung Heui Ryu, Hwan-Gue Cho, DoHoon Lee. *COCAW: Comparative Observer for Conserved Areas among Whole Genomes*

P65/ Ran Elkon, Rita Vesterman, Nira Amit, Igor Ulitsky, Gilad Mass, Idan Zohar, Dorit Sagir, Jackie Assa, Yosef Shiloh, Ron Shamir.

SPIKE: Signaling Pathways Integrated Knowledge Engine

P66/ S. A. Arul Shalom, Manoranjan Dash, Minh Tue. *GPU-Based Fast K-means Clustering of Gene Expression Profiles*

P67/ Vanishree Mallur Srinivas, Lokanath Khamari, Ruby K. Mathew, Jignesh Bhate, Jaya Iyer. *Potential Cardiovascular Disease Markers: CliPro^{TM}-Based Analysis*

P68/ Marvin N. Steijaert, Huub M.M. ten Eikelder, Anthony M.L. Liekens, Dragan Bosnacki, Peter A.J. Hilbers. *Stochastic Switching Behavior of a Bistable Auto-Phosphorylation Network*

P69/ B Bharath Bhat, Ashwin Ram B, Arathi Raghunath, Khamari Lokanath, Sadanandan Vidyendra, Jignesh Bhate, Usha Mahadevan. *Ubiquitin-Proteasome Pathway Genes and Prostate Cancer*

P70/ Meelis Kull, Jaak Vilo. *Fast Approximate Hierarchical Clustering using Similarity Heuristics and Adaptation to Time Constraints*

P71/ Gongjin Dong, Yantao Qiao, Yu Lin, Shiwei Sun, Chungong Yu, Dongbo Bu. *A Training-Set-Free Stochastic Model for Peptide Identification*

P72/ Hans-Juergen Thiesen, Peter Lorenz, Zilliang Qian, Yixue Li, Michael Kreutzer, Michael O. Glocker. *In Silico Charactization of Peptide Epitopes Recognized by Autoantibodies Present in IVIG Sample Preparations*

P73/ Niko Beerenwinkel, Nicholas Eriksson,Volker Roth, Osvaldo Zagordi. *Ultra-Deep Sequencing of Genetically Heterogeneous Samples*

P74/ Dong L. Tong, Robert Mintram. *Microarray Gene Recognition Using Multiobjetive Evolutionary Techniques*

P75/ Marcel Martin, Sven Rahmann. *A Heuristic Clustering Algorithm Using Graph Transitivity*

P76/ Sebastian Boecker, Florian Rasche. *Analysis of Metabolite Tandem Mass Spectra*

P77/ Sang Yup Lee, Jin Hwan Park, Kwang Ho Lee, Tae Yong Kim. *L-Valine Production by Systematically Engineered Escherichia coli*

P78/ Sang Yup Lee, Tae Yong Kim. *Deciphering the Evolution and Metabolism of Mannheimia succiniciproducens MBEL55E by Genome-Scale Analysis*

P79/ Duygu Tas, Kemal Kılıc, Osman Ugur Sezerman. *A New Probabilisitic Approach for Simplified Partial Digest Problem*

P80/ Elvin Coban, Kemal Kılıc, Osman Ugur Sezerman. *Constraint Programming Applied to Simplified Partial Digest Problem with Errors*

P81/ Hsin-Nan Lin, Wen-Lian Hsu. *GACOT: A Genetic Algorithm for the Physical Mapping Problem with Noisy Data*

P82/ Tobias Marschall, Sven Rahmann. *Probabilistic Arithmetic Automata and their Application to Pattern Matching Statistics*

P83/ Tae Yong Kim, Soo Yun Moon, Soon Ho Hong, Sang Yup Lee. *Metabolic Engineering of Escherichia coli for Production of Malic Acid*

P84/ Tae Yong Kim, Hyung Rok Choi, Sang Yup Lee. *Metabolic Pathway Analysis and its Optimization for Producing Succinic Acid in Mannheimia succiniciproducens MBEL55E*

P86/ Fang Rong Hsu, Wei-Chung Shia. *A Decision Support System for Cardiovascular Disease Using Bioinformatics Approach*

P87/ Sang Yup Lee, Kwang Ho Lee, Jin Hwan Park, Tae Yong Kim. *Metabolic Engineering of Escherichia coli for L-Threonine Production based on Systems Biology*

P88/ Tae Yong Kim, Hyun Uk Kim, Joon Haeng Rhee, Sang Yup Lee. *Application of Genome-Scale Metabolic Model of Vibrio vulnificus CMCP6 for In Silico Drug Targeting*

P89/ Roel G.W. Verhaak, Laura MacConaill, Carsten Russ, Jen Chen, Brian Desany, Danny A Milner Jr, Matthew Meyerson. *Pathogen Discovery by Combination of*

Computational Substraction and Pyrosequencing Technology

P90/ Thasso Griebel, Malte Brinkmeyer, Sebastian Boecker. *EPoS: A Modular Framework for Phylogenetic Analysis*

P91/ S. Avinash Kumar, S. Sundar Raman, R. Parthasarathi, V. Subramanian. *A New Triad Based Approach to Sequence Comparison of Various Types of Collagen*

P92/ Sumeet Dua, Shirin A. Lakhani, Hilary W. Thompson. *Structural Classification Using Mining of Frequent Patterns in Concave Protein Surfaces*

P93/ Ankit Rakha, Mitra Basu, Rao Kosaraju. *HLA Class I Peptides: Exploiting Positional Information for Identification and Classification*

P94/ John Thomas, Naren Ramakrishnan, Chris Bailey-Kellogg. *Protein Design by Sampling an Undirected Graphical Model of Residue Constraints*

P95/ Wei Wang, Youling Guo, Yuexian Zou, Tianrui Wu. *A Novel Algorithm for Tag SNP Selection based on Pair-Wise Linkage Disequilibrium*

P96/ Valentina Boeva, Julien Clement, Mireille Regnier, Mikhail A. Roytberg, Vsevolod J. Makeev. *Exact P-value calculation for clusters of TFBSs. Application in Computational Annotation of Regulatory Sites*

P97/ Mario Albrecht, Christoph Welsch, Francisco S. Domingues, Gabriele Mayr, Andreas Schlicker, Stefan Zeuzem, Thomas Lengauer. *Residue Interaction Networks for Analyzing Resistance Mutations in HCV Protein Structures*

P98/ Johan Rung, Ghislain Rocheleau, Alexander Mazur, Christian Dina, Constantin Polychronakos, Philippe Froguel, Rob Sladek. *A Two-Stage Genome-Wide Association Study of Type 2 Diabetes Mellitus in a French Population*

P99/ Gerard Wong, Kylie Gorringe, Ian Campbell, Izhak Haviv, Christopher Leckie, Adam Kowalczyk. *Detecting Significant Micro-Regions of DNA Aberration in High Density SNP Array Data*

P101/ Marc Delarue, Patrice Koehl. *Biomolecular Electrostatics: Beyond the Poisson-Boltzmann Centric View*

P102/ Hao Zhao, Guillaume Bourque. *Prediction and Analysis of Reliable Rearrangement Events in Mammalian Evolution*

P103/ Maryam Nikousaleh, Armin Madadkar Sobhani, Bahram Goliaei. *3D Structure Prediction of Camel Alpha-Lactalbumin*

P104/ Chi Ho Lin, Guillaume Bourque, Patrick Tan. *Comparative Analysis of Burkholderia Species Reveals an Association between Large-scale Genome Rearrangements and Fine-scale Nucleotide Variation in Prokaryotes*

P105/ Yuerong Zhu, Yuelin Zhu, Wei Xu. *EzArray: A Web-Based Highly Automated Affymetrix Expression Array Data Management and Analysis System*

P106/ Holger Froehlich, Mark Fellmann, Annemarie Poustka, Holger Sueltmann, Tim Beissbarth. Estimating Signaling Networks Through Nested Effects Models

P107/ Utz J. Pape, Martin Vingron. *Statistics for Co-Occurrence of DNA Motifs*

P108/ Mohamed Helmy, Masaru Tomita, Masa Tsuchiya, Kumar Selvarajoo. *Computational Simulations Suggest Transcription Factors AP-1 and NF-kB are Key Regulators of TLR3 Signaling*

P109/ Lee Hazelwood, John M. Hancock. *Modelling Metabolic Processes in Insulin-Secreting Pancreatic β-Cells*

P110/ Christine Steinhoff, Matteo Pardo, Martin Vingron. *Unsupervised Joint Analysis of ArrayCGH, Gene Expression Data and Supplementary Features*

P111/ H. Liu, G. Alexe, D. Juan,T. Antes, C. Delisi, L. Liou, S. Ganesan, G. Bhanot. *A*

Procedure to Identify MicroRNA Gene Targets in Human Kidney Cancer

P112/ Trupti Joshi, Chao Zhang, Ning Lin, Dong Xu. *An Integrated Probabilistic Approach for Gene Function Prediction Using Multiple Sources of High-Throughput Data*

P113/ Amit Nagal, O. P. Jangir. *Computational Studies of Lens Regeneration Under Influence of Vitamin A and its Metabolite*

P115/ Jayasree Ganugapati, Ravindra Babu Potti, Ashok Chakravarthy. *In Silico Modeling of Pesticidal Crystal-Like Protein Cry16Aa from Clostridium bifermentans*

P116/ Gabriela Alexe, Erhan Bilal, Nilay Sethi, Lyndsay Harris, Vasisht R. Tadigotla, Shridar Ganesan, Gyan Bhanot. *Patterns of Differential Over-Expression of the Oncogene IKBKE in HER2+ and Basal Breast Cancer*

P117/ Kok Siong Ang, Rudiyanto Gunawan. *Parameter Estimation of Oscillatory Systems*

P118/ Suresh K Poovathingal, Rudiyanto Gunawan, Jan Gruber, Barry Halliwell. *Aging Studies: A Stochastic Approach*

P119/ Faraaz N. K. Yusufi, Satty Ganeswara Reddy, May May Lee, Dong-Yup Lee. *GlycoVault: An Online Storage and Visualization System for Glycan Structures*

P120/ Jong Myoung Park, Hongseok Yun, Sang Yup Lee. *MFAML: Metabolic Flux Analysis Markup Language*

P121/ Jong Myoung Park, Hongseok Yun, Jeong Wook Lee, Joonwoo Jeong, Jaesung Chung, Sang Yup Lee. *EcoProDB: The Protein Database for Escherichia coli*

P122/ Jong Myoung Park, Choamun Yun, Hongseok Yun, Sunwon Park, Sang Yup Lee. *Development of an Integrative Online Tool for Modeling and Simulation of Cellular Networks*

P123/ Noel G Faux, Richard Tothill, Justin Bedo, David Bowtell, Adam Kowalczyk. *Gene Expression Profiling for the Classification of Cancers of Unknown Primary*

P124/ Janusz Dutkowski, Jerzy Tiuryn. *Inference of Protein-Protein Interactions: An Evolutionary Approach*

P125/ Xing Yi Woo, Edison T. Liu, Guillaume Bourque. *Integrative Analysis of Transcriptome and Genomic Aberration Map in Cancer*

P126/ Fang Rong Hsu, Wen Chun Lo. *Discovery of Novel Relationship Among Single Nucleotide Polymorphisms, Alternative Splicing Events and Tumor*

P127/ Arathi Raghunath, Sangjukta Kashyap, Usha Mahadevan, Jignesh Bhate, Pratap Dey. *Role of Interaction Databases in Studying Cross-Talks Between Pathways*

P128/ Fang Rong Hsu, Dung-Lin Hsieh, Shao-Peng Yeh. *AVATAR II: An Alternative Splicing Database Using Three Alignment Tool*

P131/ Lawrence Buckingham, Xin-Yi Chua, James M. Hogan, Paul Roe, Jiro Sumitomo. *Large-Scale Comparative Studies in GPFlow*

P132/ Vincent Piras, Alessandro Giuliani, Naoki Fujikawa, Masaru Tomita, Kumar Selvarajoo, Masa Tsuchiya. *Is Transcription Factors Mediated Gene Regulation Hard Wired? A Microarray-Based Statistical Estimate*

P133/ Arathi Raghunath, Pratap Dey, Usha Mahadevan, Jignesh Bhate, Sangjukta Kashyap. *Contrast Interaction Database: A Novel Approach to Study Contextual Relevance of Interactions*

P134/ Dang Hung Tran, Kenji Satou, Tu Bao Ho. *Finding MicroRNA-mRNA Modules Based on Rule Induction*

P135/ V. Jayaraj, R. Suhanya, M. Vijayasarathy, E. Rajasekaran. *Computational Studies on Role of Large Hydrophobic Residues in Proteins*

P136/ Anne Bergeron, Julia Mixtacki, Jens Stoye. *HP Distance via Double Cut and Join Distance*

P137/ Yantao Qiao, Shiwei Sun, Gongjin Dong, Yu Lin, Chungong Yu, Dongbo Bu. *A Novel Scoring Scheme to Evaluate Match of Peptide and Mass Spectrum*

P138/ Yuriy L. Orlov. *Sequence Complexity Measures for Alignment Free Genome Comparisons*

P139/ Chia-Lang Hsu, Yen-Hua Huang, Ueng-Cheng Yang. *A Gene Ontology-Based Method to Present Pathway Relations*

P140/ Armando D. Solis, S. Rackovsky. *Information-Guided Knowledge-Based Potential Functions for Protein Structure Prediction*

P141/ Rajaraman Kanagasabai, Christopher Baker. *Extraction and Grounding of Protein Mutations via Ontology-centric Knowledge Integration*

P142/ Przemyslaw Biecek, Adam Zagdanski, Rafal Kustra, Stanislaw Cebrat. *Improving Detection Performance for Gene Set Functional Enrichment and Finding Transcription Factor Binding Sites*

P144/ Brian J. Parker, Jiayu Wen, Georg F. Weiller. *Structural Strand Asymmetry for Transcription Orientation Prediction in Unaligned ncRNA Sequences*

P145/ Xin Li, Jing Li. *Zero Recombinant Haplotype Inference with Missing Data*

P146/ Jia-Ming Chang, Emily Chia-Yu Su, Allan Lo, Hua-Sheng Chiu, Ting-Yi Sung, Wen-Lian Hsu. *A Document Classification Strategy to Predict Protein Subcellular Localization Using Sequence Motifs and Evolutionary Information*

P147/ A. Ng, J.C. Rajapakse, J.G. Evans, R. Welsch. *Statistical Analysis of Macrophage Cell Morphology after Microtubule Disruption*

P148/ Reetal Pai, James Sacchettini, Thomas Ioerger. *Fragment-Based Analysis of Protein-Ligand Interactions Using Localized Stereochemical Features*

P149/ Avinash Kumar S., Namit Bharija. QSAR Studies of Anti Influenza Neuraminidase Inhibitors [Oseltamivir]

P150/ Sandro Andreotti, Juergen Kleffe, Paul Wrede. *De Novo Design of Peptides: Potential Vaccines against the Influenza A Virus*

P151/ Merlin Veronika, James G. Evans, Paul Matsudaira, Roy E. Welsch, Jagath C. Rajapakse. *Size-Specific and Brightness-Weighted Cell Tracking in 2D images*

P152/ Chi-Yong Cho, Dae-Soo Kim, Jae-Won Huh, Heui-Soo Kim, DoHoon Lee, Hwan-Gue Cho. *EVOG: Evolution Visualizer for Overlapping Genes*

P153/ Meng-Chang Hsiao, Chien-Ming Chen, Tun-Wen Pai, Wen-Shyong Tzou, Ron-Shan Chen. *A Study of Microsatellites Dominating Mammalian Size Variation*

P154/ Duygu Ucar, Fatih Altiparmak, Hakan Ferhatosmanoglu, Srinivasan Parthasarathy. *Investigating the Promise of Extrinsic Similarity Measures for Gene Expression Analysis*

P155/ Sharlee Climer, Alan R. Templeton, Weixiong Zhang. *A Dense Graph Model for Haplotype Inference*

P156/ Neil D. Clarke, Hock Chuan Yeo, Zhen Xuan Yeo, Ye Li. *Large-Scale Inference of Condition-Specific Regulation Using Gene Expression Data and Predicted Transcription Factor Occupancy of Promoters*

RECOMB 2008

Posters

Computer-Aided Multi-Parameter Antigen Design: Impact of Synthetic Peptide Vaccines

V. S. Gomase,[1] S. V. Gomase,[1] S.Tagore,[1] D. A. Bhiwgade,[1]
M. M. V. Baig,[2] K. V. Kale[3]

1 Introduction

The black widow, *Latrodectus spp.*, is considered the most venomous spider. Only the larger immature female and adult female spiders can bite through all layers of human skin and inject enough venom to cause a painful reaction. Alpha latrotoxin, a component of black widow spider venom (BWSV), causes neurotransmitter release at neuromuscular junctions and may act by forming cation-permeable pores in lipid membranes. Alpha-Latrotoxin depolarizes delta psi p selectively, both in the presence and absence of Ca2+. The antigenic epitopes on alpha latrotoxin *Latrodectus tredecimguttatus* (black widow) are important determinant of protection against spider venom [1, 2, 3].

2 Methodology

Antigenic epitopes of alpha latrotoxin *L. tredecimguttatus* is determined using the Gomase (2007), Hopp and Woods, Welling, Parker and Protrusion Index (Thornton) antigenicity methods [4] The MHC peptide binding of alpha latrotoxin is predicted using neural networks trained on C terminals of known epitopes. In analysis predicted MHC/peptide binding of alpha latrotoxin is a log-transformed value related to the IC50 values in nM units. MHC2Pred predicts peptide binders to MHCI and MHCII molecules from protein sequences or sequence alignments using Position Specific Scoring Matrices (PSSMs). Support Vector Machine (SVM) based method for prediction of promiscuous MHC class II binding peptides. SVM has been trained on the binary input of single amino acid sequence [5, 6, 7, 8, 9]. In addition, we predict those MHC ligands from whose C-terminal end is likely to be the result of proteosomal cleavage.

3 Results and Interpretations

In this analysis of antigenic determinant site of alpha latrotoxin protein, we got fifty nine (59) antigenic determinant sites in the sequence. The peptide segments in this region are 375-IGDWRDGREVRYAV-388, 524-KKGYTPIHVAADS-536, 592-KDGFTPLHYAIRG-604, 832-PIHGAAMTGLLDV-844, 965-RDECP-NEECAISHFAVCDAVQ-985, 1211-LQTNQISNFIDRK-1223, 1289-LSITEKFEDVLNSL-1302, 1383-HL-FGESCLHSDGILTK-1398, of protein called the antigenic determinant or the epitope is sufficient for eliciting the desired immune response; see Table 1. The average propensity for alpha latrotoxin protein is found to be above 1.0208. Furthermore, this region forms beta sheet. Thus beta sheet shows high antigenic response than helical region of this peptide. Regions preferably select peptides lying in MHC-Cls1-EPTHLA-A2.1-RM, GEN-T-CELL-EP motifs regions. According to Kyte-Doolittle, Hopp-Woods plot we can predict that these peptides are hydrophobic in nature. Predicted antigenic epitope is choosing peptides that are in the N-terminal region of the alpha latrotoxin. Because the N- and C- terminal regions of proteins are usually solvent accessible and unstructured, antibodies against those regions are also likely to recognize the native protein. These regions are antigenic in nature and form antibodies. These MHC Class peptide segments are from a set of aligned peptides known to bind to a given major histocompatibility complex (MHC) molecule as the predictor of MHC-peptide binding. Binding ability prediction of antigen peptides to MHC class molecules is important in vaccine development. The method integrates prediction of peptide MHC class binding, proteasomal C terminal cleavage, and TAP transport efficiency of alpha latrotoxin protein.

[1]Dept of Bioinformatics, Dr. D. Y. Patil Instituite for Biotechnology and Bioinformatics, Plot No-50, Sector-15, CBD Belapur, Navi Mumbai 400614, India. Email: virusgene1@yahoo.co.in

[2]Dept of Biotechnology, Teshwant Mahavidyalaya, VIP Road, Nanded, MS 431602, India

[3]Dept of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, MS 431004, India

4 Conclusions

Reactions to the alpha latrotoxin of the *L. tredecimguttatus* are causing human health hazard. The venom of the black widow spider *Latrodectus mactans* contains a family of related neurotoxins, known as latrotoxins, which cause dramatic stimulation of exocytosis at synapses and from endocrine cells. Antigenic peptides or fragments from alpha latrotoxin involved multiple antigenic components to direct and empower the immune system to protect the host from allergic infection. Antigenic determinants sites shows highly antigenic nature and form beta sheets in secondary structure; also showing hydrophobic characteristics. MHC molecules of alpha latrotoxin having cell surface peptides, which take active part in host immune reactions and involvement of MHC class-I and II in response to almost all antigens. Predicted MHC binding regions acts like red flags for antigen specific and generate immune response against the parent antigen. These small peptides fragments of antigen can induce immune response against whole antigen. Antigenic epitopes of alpha latrotoxin protein are important antigenic determinants against the allergic reactions. This theme is implemented in designing subunit and synthetic peptide vaccines. The antigenicity analysis method is allows potential drug targets to identify active sites, which form antibodies against alpha latrotoxin.

No	Start Pos	End Pos	Peptide	Length
8	206	215	TPTDDSLQAP	10
11	299	308	TSNNEGLLDR	10
14	358	367	TPENFAQISF	10
15	375	388	IGDWRDGREVRYAV	14
16	406	416	VSVREKACPTL	11
20	465	476	PDSAVGFKEFTK	12
25	524	536	KKGYTPIHVAADS	13
28	592	604	KDGFTPLHYAIRG	13
34	798	809	TPLHLATFKGKS	12
36	832	844	PIHGAAMTGLLDV	13
37	870	880	AAQNSHIDVIK	11
43	965	985	RDECPNEECAISHFAVCDAVQ	21
44	1046	1056	NGHFTVVQYLV	11
46	1075	1086	KAITKNHLQVVQ	12
47	1110	1120	VAENALDIAEY	11
48	1144	1155	LAVYYKNLQMIK	12
51	1211	1223	LQTNQISNFIDRK	13
55	1289	1302	LSITEKFEDVLNSL	14
58	1364	1372	SVSLPEVTD	9
59	1383	1398	HLFGESCLHSDGILTK	16

Table 1: Antigenic epitopes of alpha latrotoxin protein.

References

[1] Gomase, V.S., Kale, K.V., Sherkhane, A.S. and Narshinge, A.P. 2006. Insilico prediction of antigenic epitope and MHC, T-cell motifs regions of alpha latrocrustotoxin from latrodectus mactans. In: *Indo-Australian Symposium on Pharmacogenomics*, Manipal, India, 10 March 2006.

[2] Gomase, V.S. 2006. Prediction of Antigenic Epitopes of Neurotoxin Bmbktx1 from Mesobuthus martensii. *Current Drug Discovery Technologies*, Vol. 3, No. 3, 225–229.

[3] Rosenthal, L. and Meldolesi, J. 1989. Alpha-latrotoxin and related toxins. *Pharmacol. Ther.*, Vol. 42, 115–134.

[4] Gomase, V.S. and Changbhale S.S. 2007. Antigenicity Prediction in Melittin: Possibilities of in Drug Development from Apis dorsata. *Current Proteomics*, Vol. 4, No. 2, 107–114.

[5] Bhasin, M. and Raghava,G.P. 2005. Pcleavage: An SVM-based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. *Nucleic Acids Res.*, Vol. 33, W202–207.

[6] Reche, P.A. et al., 2002. Prediction of MHC Class I Binding Peptides Using Profile Motifs. *Human Immunology*, Vol. 63, 701–709.

[7] Buus, S. et al., 2003. Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens*, Vol. 62, 378–384.

[8] Nielsen, M. et al., 2003. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.*, Vol. 12, 1007–1017.

[9] Gomase, V.S. and Chande, A.G. 2005. Prediction of Antigenicity of Neurotoxin M9 for Possibilities in Drug Developments. *Bioinformatics India Journal*, Vol. 3, No. 4.

Prediction of MHC Binder for Fragment-Based Viral Peptide Vaccines from Cabbage Leaf Curl Virus

V. S. Gomase,[1] S. V. Gomase,[1] S. Tagore,[1] D. A. Bhiwgade,[1]
M. M. V. Baig,[2] K. V. Kale[3]

1 Introduction

Cabbage leaf curl viral peptides are most suitable for subunit vaccine development because with single epitope, the immune response can be generated in large population. This approach is based on the phenomenon of cross-protection, whereby a plant infected with a mild strain of virus is protected against a more severe strain of the same virus. The phenotype of the resistant transgenic plants includes fewer centers of initial virus infection, a delay in symptom development, and low virus accumulation. Pathogenicity protein from Cabbage leaf curl virus is necessary for its production in or on all food commodities. An exemption from the requirement of a tolerance is established for residues of the biological plant pesticide [1].

2 Methodology

Antigenic epitopes of pathogenicity protein from Cabbage leaf curl virus is determined using the Gomase (2007), Hopp and Woods, Welling, Parker and Protrusion Index (Thornton) antigenicity methods [2, 3, 4]. The MHC peptide binding of pathogenicity proteins is predicted using neural networks trained on C terminals of known epitopes. In analysis predicted MHC/peptide binding of pathogenicity proteins is a log-transformed value related to the IC50 values in nM units. MHC2Pred predicts peptide binders to MHCI and MHCII molecules from protein sequences or sequence alignments using Position Specific Scoring Matrices (PSSMs). Support Vector Machine (SVM) based method for prediction of promiscuous MHC class II binding peptides. SVM has been trained on the binary input of single amino acid sequence [5, 6, 7, 8, 9]. In addition, we predict those MHC ligands from whose C-terminal end is likely to be the result of proteosomal cleavage.

3 Results and Interpretations

RankPep server predicts binding of peptides to a number of different alleles using Position Specific Scoring Matrix (PSSM). A pathogenicity protein sequence is 295 residues long, having antigenic MHC binding peptides. MHC molecules are cell surface glycoproteins, which take active part in host immune reactions and involvement of MHC class-I and MHC II in response to almost all antigens. PSSM-based server predict the peptide binders to MHCI molecules of pathogenicity protein sequence are as 11mer_H2_Db, 10mer_H2_Db, 9mer_H2_Db, 8mer_H2_Db and also peptide binders to MHCII molecules of pathogenicity protein sequence as I_Ab.p, I_Ag7.p, I_Ad.p, analysis found antigenic epitopes region in putative pathogenicity protein (Table 1). We also found the SVM-based MHCII-IAb peptide regions 109-FSLKDPIPW, 153-PFRAPTVKI, 194-IGLTGPGPI, 139GKLKLSTAK (optimal score is 0.952); MHCII-IAd peptide regions 248-GDSASQAGL, 223-TESEVENAL, 262-TITMSVAQL, 10-NAFNYIESH (optimal score is 0.804); MHCII-IAg7 peptide regions 3-SQLANAPNA, 174-SHVDYGRWE, 36-PSTAAQFTA, 248-GDSASQAGL (optimal score is 1.744); and MHCII- RT1.B peptide regions 249-DSASQAGLQ, 39-AAQFTARLN, 21-EYQLSHDLT, 37-STAAQFTAR (optimal score is 1.361) which represented predicted binders from viral pathogenicity protein. The predicted binding affinity is normalized by the 1% fractil.

[1]Dept of Bioinformatics, Dr. D. Y. Patil Instituite for Biotechnology and Bioinformatics, Plot No-50, Sector-15, CBD Belapur, Navi Mumbai 400614, India. Email: virusgene1@yahoo.co.in
[2]Dept of Biotechnology, Teshwant Mahavidyalaya, VIP Road, Nanded, MS 431602, India
[3]Dept of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, MS 431004, India

We describe an improved method for predicting linear epitopes (Table 1). The region of maximal hydrophilicity is likely to be an antigenic site, having hydrophobic characteristics, because terminal regions of pathogenicity protein is solvent accessible and unstructured, antibodies against those regions are also likely to recognize the native protein. It was shown that a pathogenicity protein is hydrophobic in nature and contains segments of low complexity and high-predicted flexibility. Predicted antigenic fragments can bind to MHC molecule is the first bottlenecks in vaccine design.

4 Conclusions

A pathogenicity proteins from Cabbage leaf curl virus peptide nonamers are from a set of aligned peptides known to bind to a given major histocompatibility complex (MHC) molecule as the predictor of MHC-peptide binding. MHCII molecules bind peptides in similar yet different modes and alignments of MHCII-ligands were obtained to be consistent with the binding mode of the peptides to their MHC class, this means the increase in affinity of MHC binding peptides may result in enhancement of immunogenicity of viral pathogenicity protein. These predicted of pathogenicity protein antigenic peptides to MHC class molecules are important in vaccine development from Cabbage leaf curl virus.

MHC ALLELE	Rank	Sequence	Residue No.	Peptide Score
I-Ab	1	FSLKDPIPW	109	0.952
I-Ab	2	PFRAPTVKI	153	0.921
I-Ab	3	IGLTGPGPI	194	0.885
I-Ab	4	GKLKLSTAK	139	0.833
I-Ad	1	GDSASQAGL	248	0.804
I-Ad	2	TESEVENAL	223	0.759
I-Ad	3	TITMSVAQL	262	0.695
I-Ad	4	NAFNYIESH	10	0.639
I-Ag7	1	SQLANAPNA	3	1.744
I-Ag7	2	SHVDYGRWE	174	1.485
I-Ag7	3	PSTAAQFTA	36	1.408
I-Ag7	4	GDSASQAGL	248	1.365
RT1.B	1	DSASQAGLQ	249	1.361
RT1.B	2	AAQFTARLN	39	1.126
RT1.B	3	EYQLSHDLT	21	0.845
RT1.B	4	STAAQFTAR	37	0.813

Table 1: Predicted promiscuous MHC class II binding peptides from pathogenicity protein.

References

[1] Gomase, V. S. and Kale, K.V. 2007. Binding affinity prediction of Cabbage leaf curl virus pathogenicity protein for TAP transporter. In: *National Symposium on Genomics, Proteomics and Bioinformatics*, 9–10 Feb 2007, Oamanabad, India.

[2] Gomase, V.S., Kale, K.V., Chikhale, N.J., and Changbhale, S.S. 2007. Prediction of MHC Binding Peptides and Epitopes from Alfalfa mosaic virus. *Curr. Drug Discov. Technol.*, 4(2):117-1215.

[3] Gomase, V. S., Kale, K.V., Jyotiraj, A. and Vasanthi, R. 2007. Identification of MHC ligands from alfalfa mosaic virus, CTDDR-2007, *Medicinal Chemistry Research*, 15(1/6), Page 160.

[4] Gomase, V. S., Kale, K.V., Dede, P.V., Patil, S.Y. and Patil, S.S. 2007. ANN-based prediction of MHC alleles in vaccine design. In: *International Conference on Intelligent Systems & Networks* (IISN-2007), Jagadhri-135003, India, 23–25 Feb 2007, pages 223–227.

[5] Gomase, V. S., Tandale, J.P., Patil, S. A. and Kale, K.V. 2006. Automatic modeling of protein 3D structure Nucleoplasmin-like viral coat protein from Cucumber mosaic virus. In: *14th International Conference on Advance Computing & Communication*, NIT, Surathkal, 20–23 Dec 2006, pages 614–615.

[6] Bhasin, M. and Raghava, G.P. 2005. Pcleavage: An SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. *Nucleic Acids Res.*, 33:W202–207.

[7] Reche, P.A. et al., 2002. Prediction of MHC Class I Binding Peptides Using Profile Motifs. *Human Immunology*, 63:701–709.

[8] Buus, S. et al., 2003. Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens*, 62:378–384.

[9] Nielsen, M. et al., 2003. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.*, 12:1007–1017.

Plant DNA Fingerprint Software

Madhu Bala Priyadarshi[1]

1 Introduction

DNA fingerprinting is an approach to accurately identify crop varieties or genotypes. This technology is useful in cases involving unauthorized use of varieties, and has given insight to breeders to view distinctly between heterozygous and homozygous varieties of plant. Also, uniqueness of DNA fingerprinting helped to legally protect new varieties of plants and animals, whether they are developed by genetic engineering, tissue culture or traditional method. By this time, enormous amount of DNA fingerprint data has been stored in huge databases. These databases are storage repository for end-users. According to IPR, there is need for a powerful tool to provide evidence for either difference or similarity between two samples of particular crop. In addition to that, biologists require simple and powerful tools to manipulate data in infinite combination. There are number of statistical software available in Internet on public domain. Unfortunately, they are either difficult to use or very expensive. With reference to this, very often statistical packages such as SAS (SAS Institute, 1985) or R (R Development Core Team, 2003) require some programming to perform desired statistical analysis. This represents seldom a problem for scientists, and give troubles to technicians or students, with low statistic and computer background. It becomes therefore relevant to develop cheap, easily accessible and user-friendly specialised software, aimed to store and analyze data at common interface.

2 Result

In order to store and analyze profile tables of crops fingerprinted at National Research Centre on DNA Fingerprinting, NBPGR, a software entitled "Crop DNA Fingerprint Database" (Fig. 1) is developed using Visual Basic environment at front end and MS Access at back end. It is an interactive software that stores and retrieves information according to the choice of user and performs data analysis. DNA fingerprint database is designed to store and analyze profile tables of crops fingerprinted. Software is dedicated to store all necessary information regarding varieties and primers in profile tables. In addition to that, it performs some of the important statistical analyses. Module for Jaccards, Dice and simple matching coefficient analysis of the software helps to know whether two varieties are different or similar. It also helps to know the extent of similarity between varieties. Comparison may be done on one to one or one to many varieties. In order to find best informative primer modules of polymorphic information, content and average number of bands per cultivars analyses is used. Genetic relationship among different primers is found by using gene diversity and resolving power analyses. Module of barcode generation develops band map for all primers in a particular profile table. The Help module is developed to provide working assistance to users. Facility had been developed to upload data directly from MS Excel worksheet to database. The search menu is developed to search crops, techniques, primers and varieties. Different types of reports were developed for different types of analyses. Step-by-step calculation report for all types of statistical analyses is also generated for convenience of the researchers/users.

3 Conclusions

"Crop DNA Fingerprint Database" is a user friendly window-based computer package for storing and analysing profile data of crop varieties and genetic stocks. The package provides windows graphical user interface that makes software more accessible for the casual computer user and more convenient for the experienced computer user. Simple menus and dialog box selections enable user to perform statistical

[1]National Research Center on DNA Fingerprinting, National Bureau of Plant Genetic Resources, New Delhi 110012, India. Email: madhu74_nbpgr@yahoo.com

analysis and produce scientifically sound report, thereby assisting user in analysing the profile data using computational tools. It will be very important tool for Scientists, Researchers, Plant Breeders and persons involved in DNA fingerprinting of crops. It would provide an interface where DNA profile can be stored and analyzed.

Acknowledgment. I thank my colleagues at National Research Centre on DNA Fingerprinting (NRCDF), who were involved in contributing their efforts for development of Crop DNA Fingerprint Database software.

Figure 1: Startup screen of Crop DNA Fingerprint Database.

References

[1] Bhat, K.V. 2001. Molecular Data Analysis. In: *NRC on DNA Fingerprinting Training Manual on Techniques for Plant DNA Fingerprinting*, 19–28 November, 2001; *NRC on DNA Fingerprinting*, National Bureau of Plant Genetic Resources, New Delhi 12, pages 46–58.

[2] Bhattacharya, E., Dandin, S.B., and Ranade, S.A. 2005. Single primer amplification reaction methods reveal exotic and indigenous mulberry varieties and similarly diverse. *Journal of Biosciences*, 30:669–677.

[3] Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32:314–331.

[4] Kirst, M., Cordeiro, C.M., Rezende, G. D. S. P., and Grattapaglia, D. Power of Microsatellite Markers for Fingerprinting and Parentage Analysis in Eucalyptus grandis Breeding Populations. *Journal of Heredity*, 96(2):161-166.

[5] Nagaraju, J., Reddy, K. D., Nagaraja, G.M., and Sethuraman, B.N. 2001. Comparison of multilocus RFLPs and PCR-based marker systems for genetic analysis of the silkworm, Bombyx mori. *Heredity*, 86(5):588–597.

[6] Paetkau, D., Calvert, W., Stirling, I., and Strobeck, C., 1995. Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*, 4:347–354.

Platform for Integrated and Accessible Bioinformatics

David Wood,[1] Mhairi Marshall,[2] Shuzhi Cai,[3] Matthew Bryant,[4] David Hansen,[5] and Dominique Gorse[6]

Integration and simultaneous access to the many thousands of disparate public data sources and tools represent a significant bioinformatics challenge. Providing a flexible mechanism for running complex queries using these resources is crucial to uncovering knowledge. To meet this challenge we are building an integrated bioinformatics platform using Sequence Retrieval System (SRS) [1] and Storage Resource Broker (SRB) [2] as the core engines, and a web services API for ease of use.

SRS automatically downloads and indexes databanks and provides access to embedded tools. Using internal SRS data entities, linked information from multiple data sources can be queried and retrieved in single actions. SRB is middleware that provides a uniform interface for connecting to distributed data resources based on their attributes rather than physical locations. We are developing a bioinformatics web services API called Cowrie, which interfaces with SRS and SRB for data and tool integration, and provides accessible data services for targeted bioinformatics tasks. Based on Cowrie, customised utility, workflow and web applications can be rapidly developed using any technology supporting web services, thus facilitating code reuse and consolidating development effort. Such applications have at their disposal the vast array of up-to-date data and tool services maintained by QFAB.

Additional to the SRS databanks, the Australian mirror for the UCSC genome browser has been implemented at the Institute for Molecular Bioscience and managed by Queensland Facility for Advanced Bioinformatics (QFAB), while a mirror of the Ensembl genome browser is hosted at Griffith University. Authenticated users who logon to SRB are able to access and query both genome browsers and SRS files under a single file hierarchy even though the data is distributed across several storage systems. All these up-to-date data are made available for download and for SQL queries through SRB and Cowrie web services.

This poster presents the hardware and software components of this platform and illustrates how this architecture is beneficial in a high data and service bioinformatics environment.

References

[1] Sequence Retrieval System (SRS), `http://www.biowisdom.com`.

[2] Storage Resource Broker (SRB), `http://www.sdsc.edu/srb/index.php/Main_Page`.

[1]ARC Centre of Excellence in Bioinformatics, The University of Queensland, Australia. E-mail: `d.wood@qfab.org`

[2]ARC Centre of Excellence in Bioinformatics, The University of Queensland, Australia. E-mail: `m.marshall@qfab.org`

[3]ARC Special Research Centre for Functional and Applied Genomics, The University of Queensland, Australia. E-mail: `s.cai@imb.uq.edu.au`

[4]Institute for Molecular Bioscience, The University of Queensland, Australia. E-mail: `m.bryant@imb.uq.edu.au`

[5]eHealth Research Centre, CSIRO, Australia. E-mail: `David.Hansen@csiro.au`

[6]Queensland Facility for Advanced Bioinformatics, The University of Queensland, Australia. E-mail: `d.gorse@qfab.org`

DiBayes: A SNP Detection Algorithm for Next-Generation Dibase Sequencing

Susan Tang, Fiona C. L. Hyland, Tomas C. Wessel, Jon Sorenson, Heather Peckham, and Francisco M. De La Vega[1]

With the advent of next-generation sequencing by ligation, there is a need to design algorithms that can establish variations between the sequenced genome and reference sequence. Because the SOLiD^TM System uses a novel 2 base color-encoding scheme to better differentiate true sequence differences from error, data is produced in the form of color calls. We have developed algorithms for SNP detection on SOLiD sequencing data. First, each genome position is evaluated to collect preliminary evidence for heterozygosity. Subsequently, candidate heterozygous positions are passed to downstream SNP detection algorithms. We developed a Bayesian algorithm that formally incorporates prior probabilities of heterozygosity, error, and GC content. Its time-accuracy profile makes it ideal for low coverage reads. For higher coverage reads, we use a fast and accurate frequentist statistical method for SNP detection. Both methods use an error model which incorporates all known sources of error, including quality values of color calls.

We evaluated the accuracy of our algorithms with sequence data from a haploid organism (S. suis), using real reads and their respective quality values. Heterozygotes were simulated at every 10th genome position, with allele ratios of 50:50, 70:30 or 90:10. At 70:30 ratio and stringent filtering, our algorithms can detect heterozygotes with a sensitivity of 98.7% and false positive rate of 1.5 x 10^-5 at $>$ 15x coverage. For positions with 6–15x coverage, we are able to detect heterozygotes with a sensitivity of 63.2% and false positive rate of 8.2 x 10^-5. The low dibase error rate of this next-generation sequencing platform makes it particularly suitable for SNP detection at low coverage and with low false positive rates.

[1]Applied Biosystems, Foster City, CA 94404 and Beverly, MA 01915, USA.

GPS 2.0: Prediction of Kinase-Specific Phosphorylation Sites in Hierarchy

Yu Xue,[1] Jian Ren, Longping Wen, and Xuebiao Yao

Identification of phosphorylation sites with their cognate protein kinases (PKs) is the foundation for understanding the functional dynamics and plasticity of various cellular processes. Although nearly 10 kinase-specific predictors were developed, numerous PKs were casually classified into sub-groups without a standard rule. And for large-scale predictions, the false positive rate (FPR) was also never addressed. In this work, we updated our previous GPS (Group-based Phosphorylation Scoring method, ver 1.10) into a new generation of GPS software (Group-based Prediction System, ver 2.0) for predicting kinase-specific phosphorylation sites in hierarchy. We adopted a PK classification established by Manning et al. as the standard rule to cluster the human PKs into a hierarchical structure with four levels, including group, family, subfamily and single PK [1]. The training data was taken from Phospho.ELM 6.0 [2] and the modified version of GPS algorithm [3, 4] was employed. Also, we defined a simple rule to calculate the theoretically maximal FPRs. Three cut-offs of high, medium and low thresholds were established with FPRs of 2%, 6% and 10% for serine/threonine kinases, and 4%, 9%, and 15% for tyrosine kinases, respectively. The performance and robustness of the prediction system were extensively evaluated by self-consistency, leave-one-out validation and 4-, 6-, 8-, 10-fold cross-validations. Compared with existing tools, GPS 2.0 carried greater computational power with superior performance (Table 1). The GPS 2.0 was implemented in JAVA and could predict kinasespecific phosphorylation sites for 408 PKs in human. Moreover, we used GPS 2.0 directly to perform a large-scale prediction of more than 13,000 mammalian phosphorylation sites and achieved highly satisfying results. In addition, we provided a proteome-wide prediction for Aurora-B specific substrates including protein-protein interaction information. The GPS 2.0 software is freely available at: http://bioinformatics.lcd-ustc.org/gps2. A snapshot of GPS 2.0 interface is shown in Figure 1.

References

[1] Manning G, Whyte DB, Martinez R, Hunter T, and Sudarsanam S. 2002. The protein kinase complement of the human genome. *Science*, 298:1912–1934.

[2] Diella F, Cameron S, Gemund C, Linding R, Via A, Kuster B, Sicheritz-Ponten T, Blom N, and Gibson TJ. 2004. Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, 5:79.

[3] Xue Y, Zhou F, Zhu M, Ahmed K, Chen G, and Yao X. 2005. GPS: A comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Res*, 33: W184–187.

[4] Zhou FF, Xue Y, Chen GL, and Yao X. 2004. GPS: A novel group-based phosphorylation predicting and scoring method. *Biochem Biophys Res Commun*, 325:1443–1448.

[1]Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science & Technology of China, Hefei, Anhui 230027, China. E-mail: xueyu@ustc.edu.cn

Predictor	Threshold	PKA		ATM		CDC2		Src	
		Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp
ScanSite	low	69.14	95.02	54.55	93.67	73.08	95.13	28.68	95.28
	medium	42.43	99.17	27.27	98.57	**29.23**	**99.26**	11.76	99.37
	high	**16.91**	**99.91**	18.18	99.70	**8.46**	**99.84**	3.68	99.94
KinasePhos 1.0	90%	85.16	90.64	89.09	83.86	72.31	86.37	47.06	89.93
	95%	80.12	94.50	87.27	89.76	63.08	92.69	38.24	93.91
	100%	58.46	98.42	81.82	96.04	**48.46**	**97.99**	25.00	97.84
KinasePhos 2.0		55.19	89.20	89.09	38.12	**13.08**	**99.72**	86.86	55.97
NetPhosK		77.74	91.18	85.45	97.60	16.92	87.79	33.09	95.39
pkaPS		89.61	90.81						
GPS 2.0		83.09	95.04	100.00	94.03	77.96	95.16	54.02	95.34
		49.26	99.17	72.73	98.62	23.12	99.26	17.24	99.43
		8.61	99.91	32.73	99.70	7.53	99.84	3.83	99.93
		89.91	90.75	/[a]	/	93.01	86.41	66.28	89.96
		84.57	94.49	/	/	89.78	92.71	57.09	94.05
		64.39	98.43	98.18	96.04	46.77	97.99	37.93	97.85
		91.69	89.25	/	/	9.14	99.72	91.19	56.03
		89.61	91.26	87.27	97.61	91.94	87.84	52.87	95.44
		89.91	90.91						

Table 1: Comparison of GPS 2.0 with previous tools, including ScanSite, KinasePhos 1.0, 2.0, NetPhosK, and pkaPS. Both the positive and negative data we tested for GPS 2.0 were submitted on these web servers. And we fixed Sp to be similar with previous tools to compare the Sn values. The performances with better values than GPS 2.0 were marked in bold. [a]Not compared, because both Sn and Sp of GPS is better.

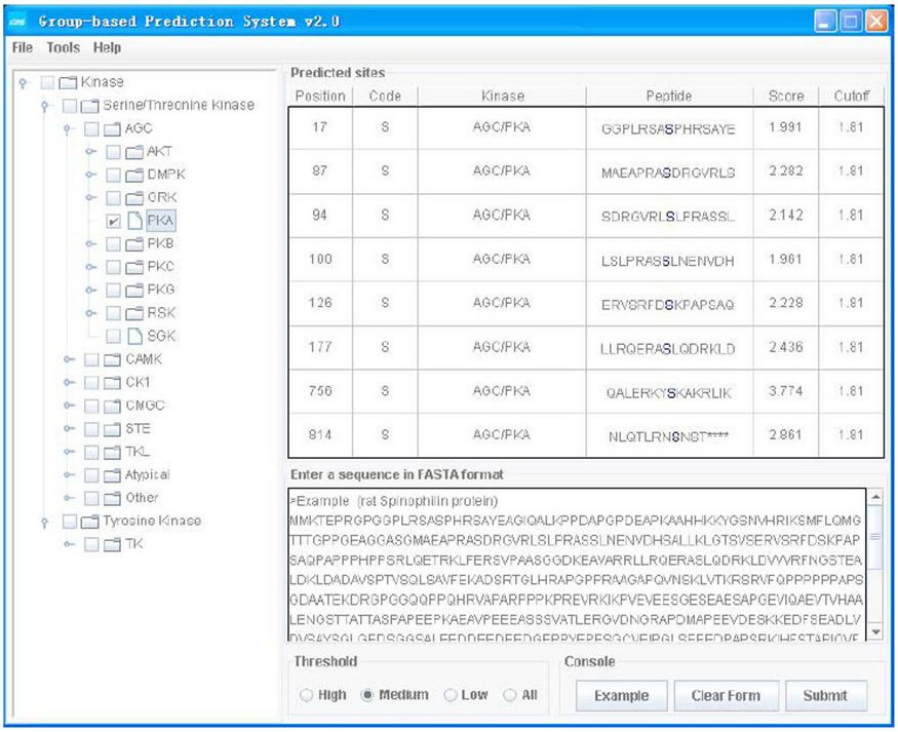


Figure 1: A screenshot of GPS 2.0 software.

Analyzing the Role of CREBBP Co-Activator in Erythroid Differentiation—a Literature Mining Approach

Jaya Iyer,[1] Arathi Raghunath,[2] Jignesh Bhate[3]

1 Introduction

All blood cells develop from pluripotent stem cells. Pluripotent stem cells differentiate into myeloid stem cells and lymphoid stem cells. Differentiation of pluripotent hematopoietic stem cells into mature circulating erythrocytes are coordinated by a set of transcription factors that are lineage and tissue specifically restricted to the hematopoietic system. Lineage specific transcription factors participate in critical protein-protein interactions in addition to binding DNA and play essential roles in red blood cell development.

The two highly related nuclear proteins CREB binding protein (CBP or CREBBP) and p300 are coactivators that possess histone acetyltransferase activity and play a central role in the integration of transcription signals involving diverse cellular processes and differentiation pathways such as hematopoiesis, embryogenesis and so on. Using data from literature mining, we have shown CBP as a transcriptional regulator of hematopoietic cell differentiation CBP in erythrocyte differentiation, a sub-pathway of myeloid stem cell differentiation.

2 Results and Discussion

An interaction database NetPro^TM was used as a source of all interactions in this study. NetPro^TM is a bimolecular interaction database involving both proteins and small molecules. The genes of interest were queried and retrieved using WebMINE (developed in-house). The interactions retrieved using a pathway/process specific query were analyzed to arrive at the network generated to explain the role of various transcription factors during the process of erythroid differentiation.

Our network analysis identifies several transcription factor nodes through which the master regulatory cofactor CBP binds and modulates the expression or activity of downstream genes involved in erythrocyte differentiation.

Using a pathway/process specific search of the published literature, the erythroid specific transcription factors regulated by CBP were identified to be GATA1, KLF1, NFE, MYB, MAFG, and SPI1 [1, 2, 3, 4, 5]. Analysis of each of the transcription factor nodal networks revealed signaling links involved in erythroid differentiation that include ZFPM1, HBZ, MAPK3, MYB, caspase (for GATA1), KIT (for MYB), HBD, HBB, HDAC (for KLF), MAFG, RAR, THR, TAF4, hemoglobin alpha (for NFE2) and FLI1 (for SPI1).

With the help of literature-mining efforts, we have been able to delineate a pivotal role for CBP as a molecular integrator in myeloid differentiation in general and a master integrator of transcription factor networks involved in erythrocyte differentiation in particular. The transcriptional network generated is an integrated representation of data collated from various literature sources. This study highlights the application of interaction databases combined with visualization in elucidating biological processes.

References

[1] Blobel, G.A, Nakajima, T., Eckner, R., Montminy, M., and Orkin, S.H. 1998. CREB-binding protein cooperates with transcription factor GATA-1 and is required for erythroid differentiation. *Proceedings of the National Academy of Sciences USA*, 95:2061–2066.

[1]Molecular Connections Pvt. Ltd, Kandala Mansions, #2/2 Kariappa Road (South Cross Road), Basavanagudi, Bangalore 560004, India. Email: jaya@molecularconnections.com

[2]Email: arathi@molecularconnections.com

[3]Email:jignesh@molecularconnections.com

[2] Cheng, X., Reginato, M.J., Andrews, N.C., and Lazar, M.A. 1997. The transcriptional integrator CREB-binding protein mediates positive cross talk between nuclear hormone receptors and the hematopoietic bZip protein p45/NF-E2. *Molecular and Cellular Biology*, 17:1407–1416.

[3] Hung, H.L., Lau, J., Kim, A.Y., Weiss, M.J., and Blobel, G.A. 1999. CREB-Binding protein acetylates hematopoietic transcription factor GATA-1 at functionally important sites. *Molecular and Cellular Biology*, 19:3496–3505.

[4] Melotti, P. and Calabretta, B. 1996. Induction of hematopoietic commitment and erythromyeloid differentiation in embryonal stem cells constitutively expressing c-myb. *Blood*, 87: 2221–2234.

[5] Starck. J., Doubeikovski, A., Sarrazin, S., Gonnet, C., Rao, G., Skoultchi, A., Godet, J., Dusanter-Fourt, I., and Morle, F. 1999. Spi-1/PU.1 is a positive regulator of the Fli-1 gene involved in inhibition of erythroid differentiation in friend erythroleukemic cell lines. *Molecular and Cellular Biology*, 19:121–135.

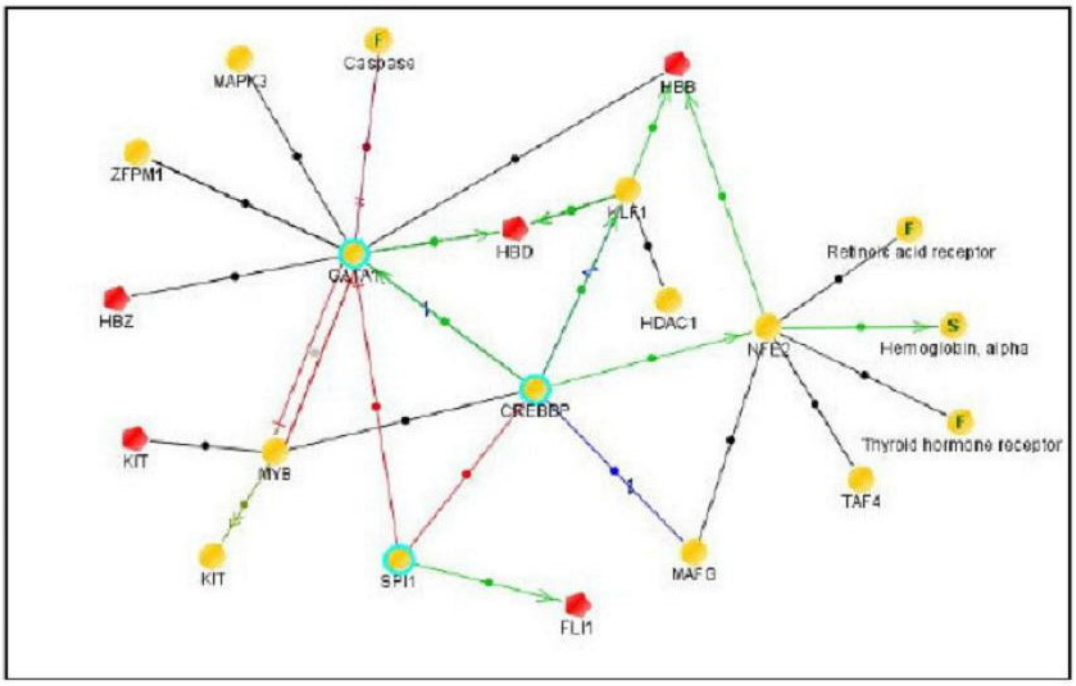


Figure 1: Interaction network highlighting the role of CREBBP in transcriptional regulation of genes involved in erythroid differentiation.

Nucleocapsid Gene Sequence Analysis and Characterization of an Indian Isolate of Avian Infectious Bronchitis Virus

Monika Koul,[1] Megha Kadam,[1] Yashpal Malik,[2]
Ashok Kumar Tiwari,[3] Jawaharlal Vegad[4]

Avian infectious bronchitis virus belongs to the family Coronaviridae. It is an enveloped virus with large positive stranded RNA genome. In the present study RNA was isolated from viral suspension and transcribed into cDNA. Poultry postmortem cases showing lesions of visceral gout were collected and infectious bronchitis virus were isolated. 1.2 kb Nucleocapsid gene of virus was amplified by RT-PCR from four clinical samples. The amplified product was cloned and the nucleotide sequence of the N gene of an Indian field isolate was determined. The Indian IBV isolate exhibited 95 per cent homology with Korean isolates and Chinese vaccine strains indicated conserved nature of N gene. Haemagglutination assay and chicken embryo inoculation was carried out for antigenic studies of the virus. The virus titre was confirmed using haemagglutination assay and IBVN2 showed the 1:2048+ titre. Propagation of virus was done by chorioallantoic method of inoculation of virus suspension in embryonated eggs. Characteristic curling and dwarfing of embryos was noticed in CAM inoculated embryonated eggs. Inoculated eggs showed teratogenic changes and deposition of urates as indication of naphropathogenic nature of virus.

[1]Biotechnology Center, Jawaharlal Nehru Krishi Vishva Vidyalaya, Jabalpur, India. Email: mkadam74@yahoo.com
[2]College of Veterinary Science and Animal Husbandry, Jawaharlal Nehru Krishi Vishva Vidyalaya, Jabalpur, India
[3]Department of Biotechnology, Indian Veterinary Research Institute, Izatnagar, Bareily, India
[4]Phoenix Poultry Diagnostics Laboratory, Jabalpur, India

Large-Scale Gene Duplication Detection in *Ciona savignyi* and *Ciona intestinalis*

Lesheng Kong[1] and Alan Christoffels[2]

1 Introduction

Species from Ciona genus have the smallest genomes of any chordate that can be manipulated experimentally [1]. As the members of the urochordates (the closest branching clade of vertebrates), Ciona species are good candidates for investigating the origins and evolutions of the chordates phylum, from which all vertebrates sprouted. In this study, we investigated the evolution of gene families corresponding to *Ciona intestinalis* and *Ciona savignyi* to better understand the factors that shape the emergence of vertebrate genomes. The protein-coding genes of *C. savignyi* were compared to genes of *C. intestinalis* and other model organisms. Gene duplication events were detected and analyzed at different levels: *C. savignyi*-specific, *C. intestinalis*-specific, Ciona-specific and Chordate-specific. Here, we reported on various over- and under-represented functional groups of duplicated genes and discuss their significance.

2 Material and Methods

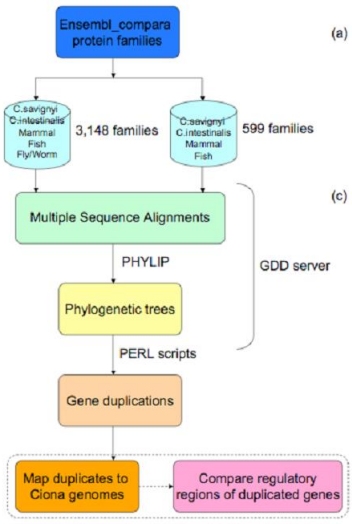


Figure 1. The flowchart of gene duplication detection. (Work in progress denoted by dashed-lines)

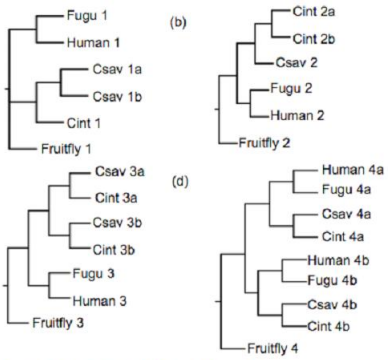


Figure 2. Example of different types of gene duplications
(a) *C. savignyi*-specific, e.g. Staphylococcal Nuclease Domain-containing Proteins (Ensembl ID: ENSF00000003054); (b) *C. intestinalis*-specific, e.g. Bifunctional Aminoacyl tRNA Synthetase (Ensembl ID: ENSF00000002044); (c) Ciona-specific, e.g. Tryptophan Dioxygenase (Ensembl ID: ENSF00000002684); and (d) Chordate-specific, e.g. Chloride Channel CLC protein (Ensembl ID: ENSF00000001538).

3 Results

Type	Families with fly/worm	Families without fly/worm
C.I.-specific	433	96
C.S.-specific	83	40
Ciona-specific	31	7
Chordate-specific	8	N.A.

Table 1: Gene duplications identified in *C. savignyi* (C.S.) and *C. intestinalis* (C.I.) genomes.

[1]Temasek Life Sciences Laboratory, National University of Singapore, Singapore 117604. Email: `lesheng@tll.org.sg`
[2]South African National Bioinformatics Institute (SANBI), University of the Western Cape, South Africa. Email: `alan@sanbi.ac.za`

GO term	Description	Type	Enrichment	P-Value
GO:0006519	amino acid and derivative metabolism	biological process	2.40	3.95e-07
GO:0009058	Biosynthesis	biological process	1.55	8.54e-05
GO:0007165	signal transduction	biological process	1.80	0.00292
GO:0004872	receptor activity	molecular function	2.55	0.00559

Table 2: Over- and under-represented GO terms in C.I.-specific duplicated genes.

GO term	Description	Type	Enrichment	P-Value
GO:0006519	amino acid and derivative metabolism	biological process	3.01	0.00661
GO:0005975	carbohydrate metabolism	biological process	2.28	0.0146
GO:0003677	DNA binding	molecular function	17.33	0.00622
GO:0007165	signal transduction	biological process	4.59	0.0146

Table 3: Over- and under-represented GO terms in C.S.-specific duplicated genes.

GO term	Description	Type	Enrichment	P-Value
GO:0005975	carbohydrate metabolism	biological process	3.13	0.00642
GO:0003682	chromatin binding	molecular function	11.7	0.00959
GO:0005515	protein binding	molecular function	3.96	0.00642
GO:0003676	nucleic acid binding	molecular function	2.23	0.0888

Table 4: Over- and under-represented GO terms in chordate-specific duplicated genes.

4 Discussions

The comparison of C.S. and C.I. genomes reveals some interesting findings: (a) There are significantly more C.I.-specific (529) duplicates than C.S.-specific (123) duplicates. The difference in gene numbers between C.S. and C.I. is likely due to significant number of C.I.-specific duplicates. (b) Overall, for Ciona duplicates, GO terms describing metabolism and nuclease activity were enriched while GO terms related to signal transduction, receptor activity and transcription were depleted. A possible explanation for these observations is: the enrichment of duplicates on metabolic functions might be associated with the rapid embryogenesis of Ciona species. Furthermore, over-expression of signal transduction genes might be deleterious to the finely-tuned cascade of enzymatic events and/or regulatory networks at work in Ciona species.

References

[1] Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., et al. 2002. The draft genome of *Ciona intestinalis*: Insights into chordate and vertebrate origins. *Science*, 298:2157–2167.

[2] Small, K.S., Brudno, M., Hill, M.M. and Sidow A. 2007. A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome. *Genome Biology*, 8:R41.

In Silico Dynamical Analysis of Cellular Systems: A Molecular Perturbation Approach

Thanneer Malai Perumal,[1] Wu Yan,[1] Rudiyanto Gunawan[1,2]

1 Introduction

Cells accomplish their myriad functions through complex regulatory networks that control cellular processes from mRNA transcription to post-translational protein activity. The term regulation implies an active dynamical response to internal and external stimuli. The complexity of a typical cellular network has been argued to provide robustness to common perturbations, but at a cost of fragility to rare mutations [5]. This complexity often limits human intuition in understanding how functional regulation is accomplished in a cell, which has motivated the use of mathematical representations to describe many cellular regulatory networks. By way of systems analysis [2, 3], such as the one presented here, these mathematical models can elucidate the mechanisms that are responsible for giving the observed cellular behavior.

In this work, we developed a novel system analysis methodology which makes use of perturbations to the molecular concentrations of each component in a cellular network. The analysis is then applied to a model of cell death regulation in Jurkat T-cells to illustrate the information that can be extracted from the results.

2 Method

The proposed dynamical analysis will focus on biological systems that can be described by common ordinary differential equations (ODEs) given by:

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}, \mathbf{p}), \quad \mathbf{x}(0) = \mathbf{x}_0$$

The novelty of this new analysis proposed here is that the sensitivities are evaluated for perturbation of the system states rather than the usual parametric change explained in [1]. The proposed sensitivity coefficient is mathematically defined as:

$$S^x_{i,j}(t,\tau) = \frac{\partial x_i(t)\ x_j(\tau)}{\partial x_j(\tau)\ x_i(t)} \text{ for } t \geq \tau,$$

which describes the relative change in the state x_i at time t due to the perturbation in the state x_j at some previous time τ. As the sensitivities are computed for perturbations in the states, the result can be validated in relatively simple experiments involving over-expression or knock-out of genes or RNA interference.

The molecular sensitivity $S_x(t,\tau)$ is an nn matrix whose rows and columns correspond to the various outputs and perturbations in the system, respectively. Each (i,j)th element can be presented in a surface contour plot. Such a plot illustrates two dynamical aspects of the perturbation-output relationship; the range(s) of time in τ that the perturbation may become significant and the range(s) of time in t that the corresponding output change appear. By analyzing either a selected perturbation (column of $S_x(t,\tau)$) or a chosen output (row of $S_x(t,\tau)$), one can obtain complementary information on the propagation of a perturbation signal through the system or the key molecules that take part in producing the observed output, respectively.

3 Application to Fas-Induced Apoptosis

We have applied the proposed analysis to a model of the cell death regulation in Figure 1 [4]. The activation of caspase-3 follows a switch-like response as shown in Figure 1 (see inset) by way of mitochondrial-independent (type-I) or mitochondria-dependent pathway (type-II). The results are shown in Figure 2

[1]Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore 117576.
[2]Email: chegr@nus.edu.sg

and 3, which illustrate the change in sensitivity levels of all the other states with respect to perturbation in FasL at time zero and the sensitivities of caspase-3 activation to the levels of the death signal FasL, caspase-8, caspase-6, and "activated" mitochondria (mitochondria after permeabilization by Bcl-2) respectively.

4 Conclusion

The analysis indicated that the cell death mainly depends on the type-II pathway as indicated by tracking the propagation of signal in Figure 2 and by the high sensitivity (darker regions) to activated mitochondria and the lack thereof to caspase-6 in Figure 3. In addition, the analysis also illustrated the timing of the key molecules in activating caspase-3: FasL is early, followed by caspase-8 and finally by mitochondria permeabilization. These findings implied that the FasL induction of cell death in this cell line primarily depends on the type-II pathway, in agreement with experiments [6].

References

[1] Aldridge BB, Haller G, Sorger PK, and Lauffenburger DA. Direct Lyapunov exponent analysis enables parametric study of transient signalling governing cell behaviour. *IEE Proc Syst Biol*, 153:425–432, 2006.

[2] Gunawan R, Cao Y, Petzold L, and Doyle FJ III. Sensitivity analysis of discrete stochastic systems. *Biophys J*, 88:2530–2540, 2005.

[3] Gunawan R and Doyle FJ III. Phase sensitivity analysis of circadian rhythm entrainment. *J Biol Rhythms*, 22:180–194, 2007.

[4] Hua F, Hautaniemi S, Yokoo R, and Lauffenburger DA. Integrated mechanistic and data-driven modelling for multivariate analysis of signalling pathways. *J R Soc Interface*, 3:515–526, 2006.

[5] Kitano H. Biological robustness. *Nat Rev Genet*, 5:826–837,2004.

[6] Scaffidi C, Fulda S, Srinivasan A, Friesen C, Li F, Tomaselli KJ, Debatin KM, Krammer PH, and Peter ME. Two CD95 (APO-1/Fas) signaling pathways. *EMBO J*, 17:1675–1687, 1998.

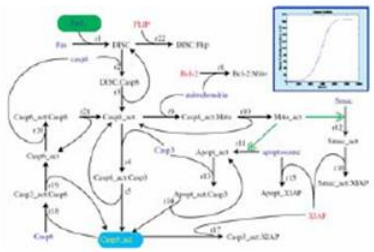


Figure 1. A Model of FasL-induced Cell Death Signaling.

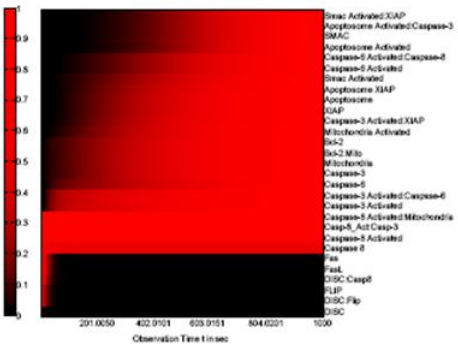


Figure 2. Change in sensitivity levels of all the other states with respect to perturbation in death signal FasL at time zero

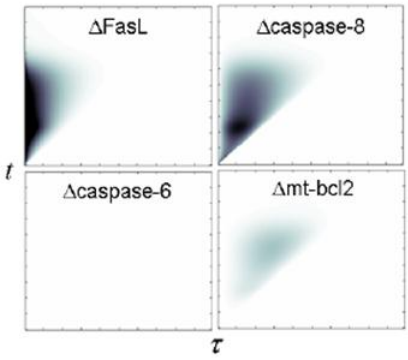


Figure 3. Molecular sensitivities of caspase-3 activation.

Design of RNA Fragments Structural Database

Marta Szachniuk,[1] Marek Blazewicz,[2]
Mariusz Popenda,[3] Ryszard W. Adamiak[4]

1 Introduction

Structural bioinformatics aims at creating a global perspective from which some unifying principles in molecular biology could be discerned. In the field of structure analysis, we observe a growing practical importance of RNA studies resulting from the recent discoveries concerning RNAi mechanism or the involvement of regulatory RNAs in cancer and other diseases [8]. This study is usually based on the results of different experiments as well as on an analysis of structural information stored in the databases. Thus, designing and creating databases to manage large amounts of biological data is a crucial bioinformatic task.

There exist several databases concerning the field of RNA structural biology. For example, Protein Data Bank [2] and Nucleic Acids Database [1] hold a collection of RNA structures, providing atom coordinates and some other structural data. The MeRNA database [6] offers classification of metal ion binding sites in RNA structures, SCOR database [4] surveys the three-dimensional RNA motifs within the PDB- and NDB- deposited structures.

Here, we describe the design of the novel relational database to store a wealth of structural data concerning RNAs and their complexes. It contains RNA sequences and secondary structures derived from PDB-deposited structures, atom coordinates of nucleotide residues, torsion angles, sugar pucker parameters, and information about base pair types. The database provides a unique opportunity to search for three-dimensional RNA fragments with primary and / or secondary structures matching a user-defined pattern. This feature can be of a great use for RNA tertiary structure prediction systems.

2 Methods

The leading idea of the project was to correlate the information about primary, secondary and tertiary structure of RNAs in such a way that whichever RNA fragment could be extracted for a comparison, at the level of the three-dimensional structure, with any other fragments having the same sequence and / or secondary structure. Moreover, we wanted to provide a possibility to search for any RNA structure satisfying user predefined conditions concerning structural parameters (e.g. torsion angles or sugar pucker parameters), sequence or secondary structure. Following these ideas, the relational database of RNA fragments has been projected. Its design allows for storing the information about primary and secondary structures of a fragment in a form of an expression composed of two mixed character chains (Figure 1). One of them corresponds to the primary structure of RNA fragment, described as a sequence of letters accepted by IUPAC-IUB codes. The second chain codes the secondary structure using dot-bracket notation [3].

The details of the three-dimensional structure are managed separately for each nucleotide residue. To query the database a user provides structural pattern which consists of the definition of the sequence and / or the secondary structure given in the dot-bracket notation. The pattern can contain inexact description of the structures. Before the search it is turned into a regular expression (c.f. Figure 1d: N=[A or C or G or U], B=[not A]). Next, a regular expression matching algorithm is performed to find the requested RNA fragment within database records. As a result, the list of RNA fragments matching the query is presented to the user. One can see the detailed information about the three-dimensional structure for each of them. The scheme of the database main part is shown in Figure 2.

[1]Institute of Computing Science, Poznan University of Technology, Poland. Email: mszachniuk@cs.put.poznan.pl
[2]Institute of Computing Science, Poznan University of Technology, Poland. Email: mblazewicz@cs.put.poznan.pl
[3]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland. Email: marpop@ibch.poznan.pl
[4]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland. Email: adamiakr@ibch.poznan.pl

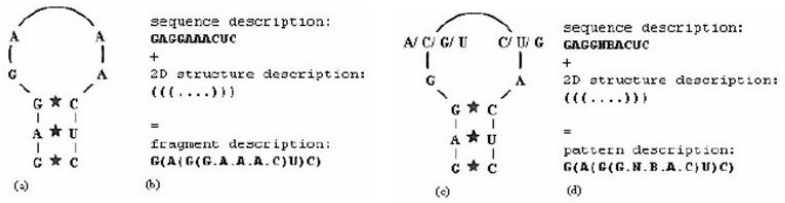


Figure 1: Example RNA fragment (a) and its encoding (b). Example search pattern (c) and its encoding (d).

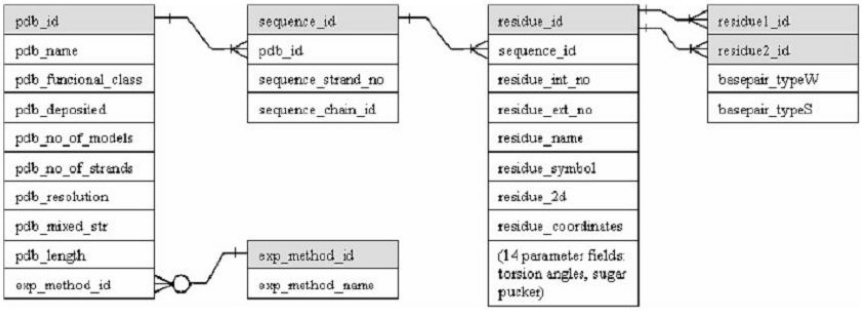


Figure 2: Scheme of the database core.

3 Results

The database has been implemented as a web-based tool RNA FRABASE, and is freely accessible at tt http://rnafrabase.ibch.poznan.pl [5]. It has been built on PostgreSQL and runs in SUSE Linux. Protein Data Bank serves as a source of selected structural data, i.e. sequences and atom coordinates of nucleotide residues. The remaining structural information stored in the database is reconstructed by our own scripts implemented in PHP and AWK. In case of secondary structure reconstruction and base pair classification, the scripts performing the coding procedures are based on RNAView software [7].

Acknowledgments. This research was supported by SP 01/04 grant of the Foundation for Polish Science and grants of the Polish Ministry of Science and Higher Education

References

[1] Berman, H.M., Westbrook, J., Feng, Z., Iype, L., Schneider, B. and Zardecki, C. 2003. The nucleic acid database. *Methods Biochem Anal.*, 44:199–216.

[2] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242.

[3] Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M. and Schuster, P. 1994. Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f. Chemie*, 125:167–188.

[4] Klosterman, P.S., Hendrix, D.K., Tamura,M., Holbrook, S.R. and Brenner, S.E. 2004. Three dimensional motifs from the SCOR, structural classification of RNA database: Extruded strands, base triples, tetraloops and U-turns. *Nucleic Acids Res.*, 32:2342–2352.

[5] Popenda, M., Blazewicz, M., Szachniuk, M. and Adamiak, R.W. 2008. RNA FRABASE version 1.0: An engine with a database to search for the three-dimensional fragments within RNA structures. *Nucleic Acids Res.*, doi:10.1093/nar/gkm786.

[6] Stefan, L.R., Zhang, R, Levitan, A.G., Hendrix, D.K., Brenner, S.E. and Holbrook, S.R. 2006. MeRNA: A database of metal ion binding sites in RNA structures. *Nucleic Acids Res.*, 34:D131–D134.

[7] Yang, H., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H. and Westhof, E. 2003. Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, 31:3450–3460.

[8] Zamore, P.D. and Haley, B. 2005. Ribo-gnome: The big world of small RNAs. *Science*, 309:1519–1524.

ASMPKS: An Analysis System for Modular Polyketide Synthases

Hongseok Tae,[1] Kiejung Park[2]

1 Introduction

Since many infectious microorganisms are acquiring tolerances for most antibiotics, the need for novel antibiotics is greatly increasing. Various methods to synthesize new antibiotics are being studied, including the approaches manipulating genes related to antibiotics biosynthesis such as polyketides. Handling known antibiotics is a more efficient approach than finding microorganisms having new kinds of antibiotics. Polyketides are secondary metabolites of many kinds of microorganisms with diverse biological functions, including pharmacological activities such as antibiotic properties. While the polyketide antibiotics are important clinical drugs, new kinds of polyketides are still being discovered. Polyketides are synthesized by serialized reactions of a set of enzymes called polyketide synthase(PKS)s [4], which coordinate the elongation of carbon skeletons by the stepwise condensation of short carbon precursors [2].

We have developed ASMPKS (an Analysis System for Modular Polyketide Synthesis) to efficiently support computational analysis of the modular PKS in genome sequences. ASMPKS operates as a web application and provides various features including visualization of polyketide structures, new PKS assembly simulation and management of the modular polyketide database.

2 Methods

ASMPKS has been developed for the overall management of modular PKS data. As it operates on the web interface to construct the database and to analyze polyketides, the extension of database is very accessible for multiple users (Fig. 1). Researcher can add and delete their data in the database easily. The database system of ASMPKS is divided into two parts. The first part, which has been constructed with published data, contains information regarding PKS genes, modules, domains and assembly. It is used to search and to align domains of protein sequences. The second part, which contains genome data of microorganisms and polyketide information related to the genomes, allows researchers to study synthesis of polyketides in a specified microorganism.

The PKS composition and the chemical structure of a polyketide in the database are displayed by the PKS navigation component, which shows the arrangements of the PKSs with their domain composition and draws the intermediate chain for a selected polyketide. The domain button has a hyperlink to homology search and multiple sequence alignment components for the analysis of domain similarity relationships. BLAST [1] is used for homology analysis between the same type domains and ClustalW [5] is used for multiple sequence alignments. And the PKS assembly component assembles a set of modules and shows the construction of an expected carbon body for a predicted polyketide.

The chemical structure of a polyketide makes its chemical activity easily understood. ASMPKS provides a PKS assembly component, which assembles a set of modules and shows the construction of an expected carbon body for a predicted polyketide. The biosynthesis of a polyketide begins by selecting a starter unit and continues by adding many extender units onto the carbon chain until a TE domain appears. As there are various kinds of starter and extender units, diverse polyketides can be constructed by their combination.

ASMPKS predicts domain information from protein sequences. Domain identification is based on the homology search method with template sequences of domains. To detect domains, BLAST is used. Template sequences that represent each domain type are formatted into the BLAST database file. To select each template sequence representing each domain type, homology scores between every pair of sequences of that type are measured, and the sequence with the highest score, which is the sum of its top 10 homology scores with other sequences, is selected. A genome analysis component is also

[1]Department of Computer Engineering, Chungnam National University, South Korea; Information Technology Institute, SmallSoft CO. LTD, Daejeon, South Korea. Email: `hstae@smallsoft.co.kr`

[2]Information Technology Institute, SmallSoft CO. LTD, Daejeon, South Korea. Email: `kjpark@smallsoft.co.kr`

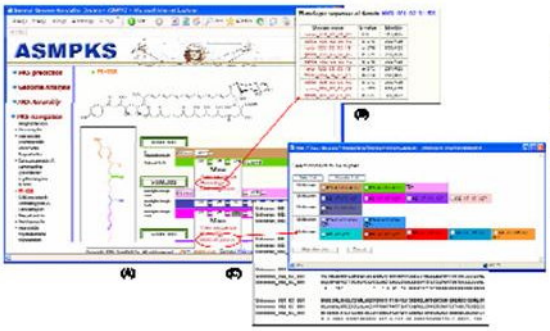


Figure 1: Information viewers for a polyketide.

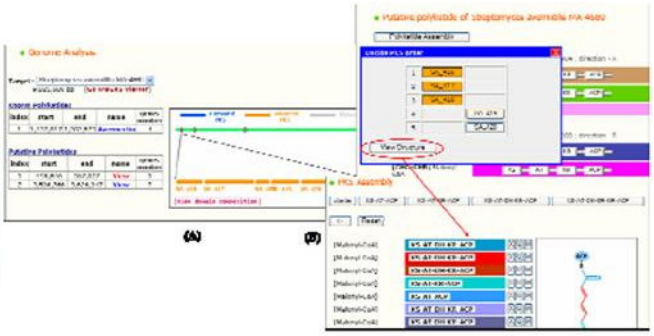


Figure 2: PKS analysis result for genome sequences.

provided. It searches microbial genome sequences for modular PKSs. It detects PKS gene clusters producing known polyketides, which are included in the database, by measuring the homology between protein sequences of an annotated genome and all PKS sequences, and predicts unknown gene clusters to produce putative polyketide candidates by identifying domains (Fig. 2). It can accept genome sequences or GenBank format data, including gene information. If genome sequences are submitted, genes are predicted by Glimmer [3], and their sequences are converted to proteins. And the automated genome wide PKS analysis, which finds known and unknown PKS gene clusters, is carried out. The result of the PKS analysis process against microbial genome sequences can be displayed in the genome browser of ASMPKS. It shows the position and composition of gene clusters of each polyketide on a genome.

3 Results and Discussion

ASMPKS has been developed for computational analysis of the modular PKS for genome sequences. It also provides overall management of information on modular PKS, including PKS database construction, new PKS assembly, and visualization of polyketide structures. It is a useful system to analyze known polyketides and to predict new polyketides. The PKS assembly and genome analysis components are especially powerful computation tools for polyketide research. As various factors are related to polyketide biosynthesis, ASMPKS can be improved through further study. ASMPKS is available from `http://gate.smallsoft.co.kr:8008/ hstae/asmpks/index.html`.

References

[1] Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[2] Cheng Y.Q., Tang G.L., Shen B. Type I polyketide synthase requiring a discrete acyltransferase for polyketide biosynthesis. *Proc. Natl. Acad. Sci. USA*, 100:3149–3154, 2003.

[3] Salzberg S.L., Delcher A.L., Kasif S., White O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, 26:544–548, 1998.

[4] Staunton J, Weissman K.J. Polyketide biosynthesis: A millennium review. *Nat. Prod. Rep.*, 18:380–416, 2001.

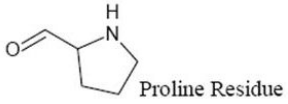
[5] Thompson J.D., Higgins D.G., Gibson T.J. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.

In-silico Analysis of HIV-1 P1 Sequence and Structure Prediction

Saurabh Shukla[1] and Alok Shekhar[1]

1 Introduction

In the genome of HIV-1, there are various sequences which have unknown function till now; p1 sequence (FLGKIWPSYKGRPGNF) is one of them, in which only 16 amino acids are there. It is assumed that two proline residues in the 7th and 13th positions are important for viral infectivity. This p1 sequence



Proline Residue

arrangement is a part of "gag protein" and not any type of structure has been determined in PDB (Protein Data Bank). So, the aim of this topic is to predict the structure of p1 sequence and then block this site after the lead designing on the basis of computational approaches of Bioinformatics. One thing is very crucial is that, p1 proline residues (position 7 and 13) are critical for replication in the HIV-1 strain HXB2-BH10. In this study we have focused on the proline rich p1-p6(Gag) C-terminus of HIV-1. Replacement of the two proline residues by leucines resulted in mutants with altered protein processing and reduced genomic RNA dimer stability that were also noninfectious.

2 Softwares and Databases Used

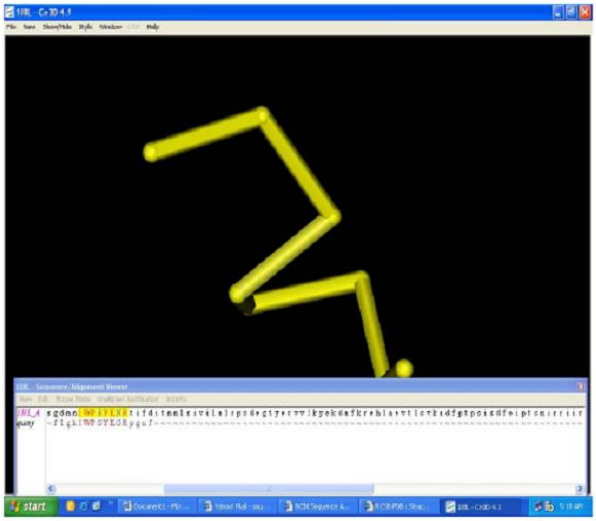
The databases and softwares used by us are given below:

1. Bioafrica Database of HIV
2. NCBI
3. MMDB Molecular Modeling Database
4. Cn3D for Visualisation
5. Raptor 3D server

[1]Yeshwant College of IT-Bioinformatics & Biotechnology, Parbhani, Maharashtra, India.

3 Results

We have successfully predicted the structure of p1 sequence by the Raptor 3D server and MMDB. According to the predicted structure, only loop part is exist in the p1 region due to the abundance of proline residues, that is why helix and sheets are absent in p1 region as we know that "Proline is the helix breaker residue".



Predicted structure of p1 sequence,(only loop part is there).

The second result is given by Raptor 3D, and it is also same, only loop part is there in p1 sequence.

```
  P1 sequence:- F L G K I W P S Y K G R P G N F
Raptor output:- C C C C C C C C C C C C C C C C (C denotes the loop region)
```

References

[1] HIV Bioafrica database.

[2] Esnouf, R., Ren, J., Ross, C., Jones, Y., Stammers, D. and Stuart, D. *Nat. Struct. Biol.* 2:303–308, 1995.

[3] JL Jain book.

[4] BD Singh book.

[5] Raptor Manual.

SiteSeek: Phosphorylation Site Predictor Using Adaptive Locality-Effective Kernel Methods and New Sequence Profiles

Paul D. Yoo,[1] Yung Shwen Ho,[2] Bing Bing Zhou,[3] Albert Y. Zomaya[4]

1 Introduction

In order to determine phosphoproteins and individual phosphorylation sites, various experimental tools have been used. However, many have indicated that in vivo or in vitro identification of phosphorylation sites is labour-intensive, time-consuming and often limited to the availability and optimisation of enzymatic reactions [1]. Several large-scale phosphoproteomic data using the mass-spectrometry approach have been collected and published [2]. These however are sill unfavourable in distinguishing the kinase-specific sites on the substrates. Due to the practical limitations and complexities of these methods, many scientists now turn to computer-based methods. These methods not only efficiently handle massive amounts of protein data but also determine phosphoprotiens and identify individual phosphorylation sites from one dimensional atomic coordinates with high precision.

Although a large number of computational methods have proved to be effective in the prediction of phosphorylation site, several important issues that can potentially degrade the performance of machine learning or statistical-based methods have been largely ignored. It has been widely recognised that the high dimensionality of protein sequence data not only causes a dynamic increase in computational complexity but also can be induced into the overfitting/generalisation problem of non-parametric methods. Hence, in this poster, we introduce a new computer-based phosphorylation site predictor, named SiteSeek which can effectively avoid the above-mentioned problems by utilising a newly developed semi-parametric machine learning model and a novel sequence profile.

2 PS-Benchmark_1 Dataset

In this study, we use a newly developed comprehensive dataset, namely PS-Benchmark_1 for the purpose of benchmarking sequence-based phosphorylation site prediction methods. PS-Benchmark_1 contains experimentally verified phosphorylation sites manually extracted from major protein sequence databases and the literature. The dataset comprises of 1,668 polypeptide chains and as shown in Table 1, the chains are categorised in four major kinase groups, namely cAMP-dependent protein kinase/protein kinase G/protein kinase C extended family (AGC), calcium/calmodulin-dependent kinase (CAMK), cyclin-dependent kinase-like kinase (CMGC) and tyrosine kinase (TK) groups. The dataset comprises of 513 AGC chains, 151 CAMK chains, 330 CMGC chains, and 216 TK chains. The dataset is non-redundant in a structural sense: each combination of topologies occurs only once per dataset. Sequences of protein chains are taken from the Protein Data Bank (PDB), Swiss-Prot, Phospho3D, Phospho.ELM and literature.

3 Compact Evolutionary and Hydrophobicity Profile

Importantly, several recent studies reported that protein hydrophobicity can be affected by the level of phosphorylation or vice versa [3]. Hydrophobicity is a very important factor in protein stability. The "hydrophobic effect" is believed to play a fundamental role in the spontaneous folding of proteins. In order to create a new profile, we use the hydrophobicity in the format of SARAH1 scale in addition to the existing sequence profile generated by PSI-BLAST. The less-discriminatory features in the sequence

[1]School of Information Technologies, University of Sydney, NSW 2006, Australia. Email: dyoo4334@it.usyd.edu.au
[2]Faculty of Medicine, University of Sydney, NSW 2006, Australia. Email: shwen_ho@wmi.usyd.edu.au
[3]School of Information Technologies, University of Sydney, NSW 2006 Australia. Email: bbz@it.usyd.edu.au
[4]Sydney Bioinformatics Centre and the Centre for Mathematical Biology, University of Sydney, Sydney, NSW 2006, Australia. Email: zomaya@it.usyd.edu.au

profile are removed by using the auto-associative network embedded in Adaptive-LEKM in order to prevent some possible problems that may be caused by the high complexity of the learner.

4 Model Proposal

The Adaptive-LEKM contains the evolutionary information represented with the local model. Its global model works as a collaborative filter that transfers the knowledge amongst the local models in formats of the hyper-parameters. The local model contains an efficient vector quantisation method. As the global model (SVM) extracts worst-case examples xi and use statistical analysis to build large margin classifiers. In Adaptive-LEKM, the original source dataset is partitioned by vector quantisation function into a set of sub-regions $P = \{S_1, S_2, \ldots, S_N\}$, then each local region is represented by the codevector c_i. The centroid vector within a cluster can be expressed as:

$$Q_i(X_m) = c_i = \frac{\sum_{X_m \in S_i} X_m}{N},\ i = 1, 2, \ldots, N$$

To construct a semi-parametric model, we substitute $Q_i(X)$ for each training sample x_i used in the SVM decision function. The basic architecture of Adaptive-LEKM is illustrated in Figure 1.

5 Results

Table 1 compares the results of SiteSeek with the consensus results of the literature. In general, SiteSeek showed about 9% better prediction accuracy than the consensus results. The experimental results of SiteSeek are written in bold and others are the consensus results of literature.

References

[1] Zanzoni A, Ausiello G, Via A, Gherardini PF, and Helmer-Citterich M. Phospho3D: A database of three-dimensional structures of protein phosphorylation sites. *Nucl. Acids Res.*, 35:D229–D231, 2007.

[2] Ballif BA, Villen J, Beausoleil SA, Schwartz D. and Gygi SP. Phosphoproteomic analysis of the developing mouse brain. *Mol. Cell. Proteomics*, 3:1093–1101, 2004.

[3] Jang HH, Kim SY, Park SK, Jeon HS,. Lee YM, Jung JH, Lee SY, Chae HB, Jung YJ, Lee KO, Lim CO, Chung WS, Bahk JD, Yun D, Cho MJ, and Lee SY. Phosphorylation and concomitant structural changes in human 2-Cys peroxiredoxin isotype I differentially regulate its peroxidase and molecular chaperone functions. *FEBS Letters*, 580(1):351–355, 2006.

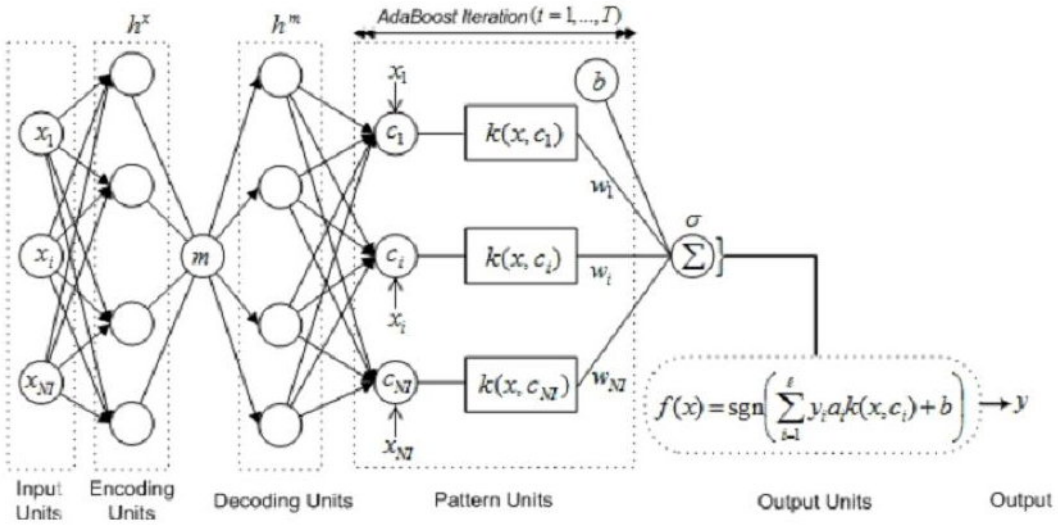


Figure 1: Adaptive-LEKM Basic Architecture.

K-Families	Accuracy (Ac)	Sensitivity (Sn)	Specificity (Sp)	Correlation-Coefficient (Cc)	Type I ER	Type II ER
CDK	**0.909**	**0.895**	**0.921**	**0.817**	**0.043**	**0.046**
	0.777	0.455	0.992	0.900		
CK2	**0.918**	**0.881**	**0.948**	**0.835**	**0.029**	**0.051**
	0.840	0.765	0.888	0.660		
PKA	**0.891**	**0.843**	**0.929**	**0.779**	**0.039**	**0.069**
	0.816	0.561	0.987	0.640		
PKC	**0.827**	**0.731**	**0.903**	**0.650**	**0.053**	**0.118**
	0.726	0.475	0.898	0.420		
Avg.	**0.886**	**0.838**	**0.925**	**0.770**	**0.041**	**0.071**
	0.790	0.564	0.941	0.655		
Var.	**0.041**	**0.074**	**0.019**	**0.083**	**0.010**	**0.032**
	0.050	0.142	0.056	0.196		

Table 1: Prediction results of Adaptive-LEKM for the four kinase families.

Exploring Protein-Protein Interactions at the Domain Level

Nandita Das, Vaibhav Navaghare, Vidyendra Sadanandan
Jignesh Bhate, Jaya Iyer[1]

1 Introduction

There is always a need to validate the detected protein-protein interactions obtained by various high-through put experiments. Vast amount of data is available on protein interactions, domain information of a protein and interacting-domains without protein in context. Given that protein-protein interactions involve physical interactions between protein domains, domain-domain interaction information can be useful for validating, annotating, and even predicting protein interactions.

2 Method and Result

We have explored protein interactions at the domain and amino acid resolution using a visualization tool developed in-house. For the same Pfam is used for protein domain information and Interdom for domain interaction information, and NetPro^{TM} for the source of protein interactions. NetPro^{TM} is an interaction database with manually curated data from articles published in PubMed and consists of protein-protein and other bimolecular interactions. NetPro^{TM} is a comprehensive database that provides all supplementary information required to understand the details of an interaction such as species, the nature of interaction (direct or indirect), the interaction terms, experimental methods, location, condition in which the interaction is taking place, domain(s) involved in an interaction, relevance of the interaction or the interacting partners to a disease condition etc. Prosite database was incorporated for residue level details of a domain. Relevant interaction details from various sources are combined and presented on a single screen.

As an example, CREB1 is taken as a query protein to comprehensively analyze its interacting partners at the domain level. The output of domain interaction details for CREB1 is represented in Figure 1. The query molecule CREB1 with its domains is shown in the center with the interacting proteins above and below. The interacting partners with single interacting domain are placed on top and the other proteins with multiple domains are placed below the query molecule CREB1. Also, the specific details like the position of the domains on the molecule, residue information etc. are clearly seen. Wherever available, the pattern details and the description from PROSITE database are also shown. Links to public databases like Pfam and PROSITE are provided.

With successful integration of interaction data points from NetPro^{TM} with the public domain databases such as Pfam, Interdom and PROSITE, the domain level interaction of the vast repository of multidomain proteins and protein complexes can be deciphered and visualized.

References

[1] Interdom Database, `http://interdom.lit.org.sg`.

[2] Pfam Database, `http://www.sanger.ac.uk/Software/Pfam`.

[3] Prosite Database, `http://au.expasy.org/prosite`.

[1]Molecular Connections Pvt. Ltd, Kandala Mansions, #2/2 Kariappa Road (South Cross Road), Basavanagudi, Bangalore 560004, India. Email: `jaya@molecularconnections.com`

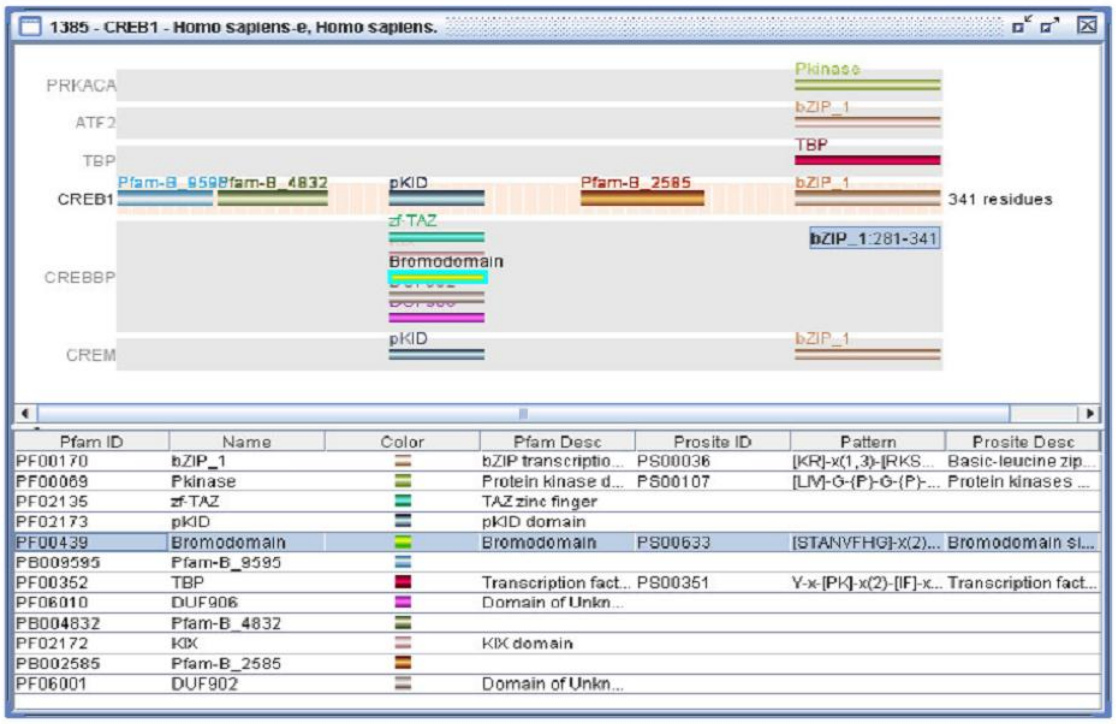


Figure 1: CREB1 with its interacting partners at the domain level.

In-silico Disease Target Screening and Evaluation with RNAi Data

Sangjukta Kashyap,[1] Nandita Das, Usha Mahadevan, Jignesh Bahate

1 Introduction

In the drug discovery industry, RNA interference (RNAi) screens are popular for drug target identification and validation. Recently RNAi therapy has become the most promising hope for the treatment of many diseases including Huntington's Disease [1]. We propose a simple in-silico method of target screening and validation based on literature evidence, which may form the basis of extensive target identification studies. The approach is based on querying data from NetPro^{TM} interaction database and NetPro^{TM} Disease module using a customized JAVA-based querying tool. NetPro^{TM} is a fully hand-curated interactome of Proteins, small molecules, DNA and RNA mined from PubMed literature with supplementary information on disease, location, experimental method, etc. The query tool is designed to provide an accessible method for screening of disease target genes, and to understand potential mechanisms of involvement of the identified targets to disease, validated by RNAi experimental data.

2 Method and Discussion

1. Select disease(s) of interest from indexed table as displayed in the query tool.
2. Query by disease(s) terms in the 'Disease' field provided in the querying tool. Result displays bimolecular interactions where the interacting molecules are involved/ associated with the queried disease. The association is displayed as 'Differential expression', 'Mutation', 'Significance', 'Therapeutic relevance', etc. Alternately the above terms may be used as filters to obtain gene association/involvement. These filters pertain to disease related molecules.
3. Query by the gene list obtained to deduce the possible mechanism of its involvement in the disease with experimental method: RNA Interference as the filter.
4. The result retrieves a bimolecular interaction experimentally proven by RNAi and the interaction result (the outcome of the interaction). The interaction result, which could be an affected cellular process or pathway, or products formed due to the interaction, may hint an association with the disease.

The tool serves as a useful testimony of concept that the ideas expressed in our approach are practical. The tool is accessible at `www.molecularconnections.com`.

References

[1] `http://www.hdlighthouse.org/research/genetherapy/updates/0057RNAi.php`.

[2] `http://expasy.org/uniprot/P24385`.

[3] `http://genomics.senescence.info/genes/entry.php?hugo=TP53`

[4] Jarskog, LF. 2006. Apoptosis in schizophrenia: Pathophysiologic and therapeutic considerations. Curr Opin Psychiatry, 19(3):307–12.

[5] `http://biol.lancs.ac.uk/gig/pages/pg/aggrecan.htm`.

[6] `http://www.wipo.int/pctdb/en/wo.jsp?WO=2004%2F029210&IA=WO2004%2F029210&DISPLAY=DESC`.

[1]Molecular Connections Pvt. Ltd., Kandala Mansions, 2/2, Kariappa Road, Basavangudi, Bangalore 560004, India. Email: `sangjukta@molecularconnections.com`

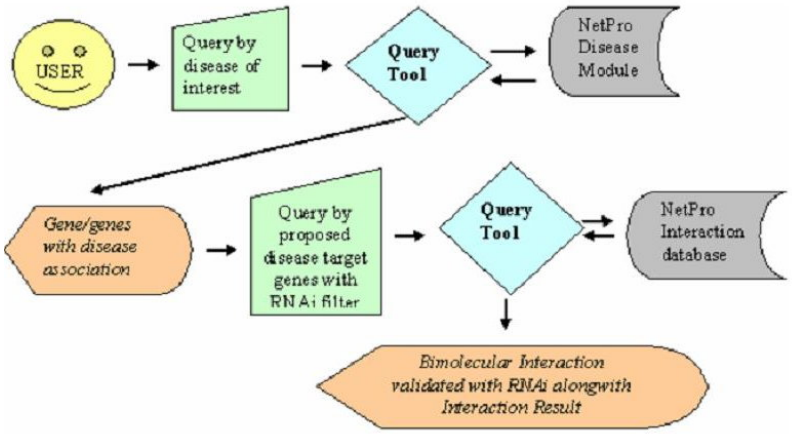


Figure 1: Overview of the query methodology.

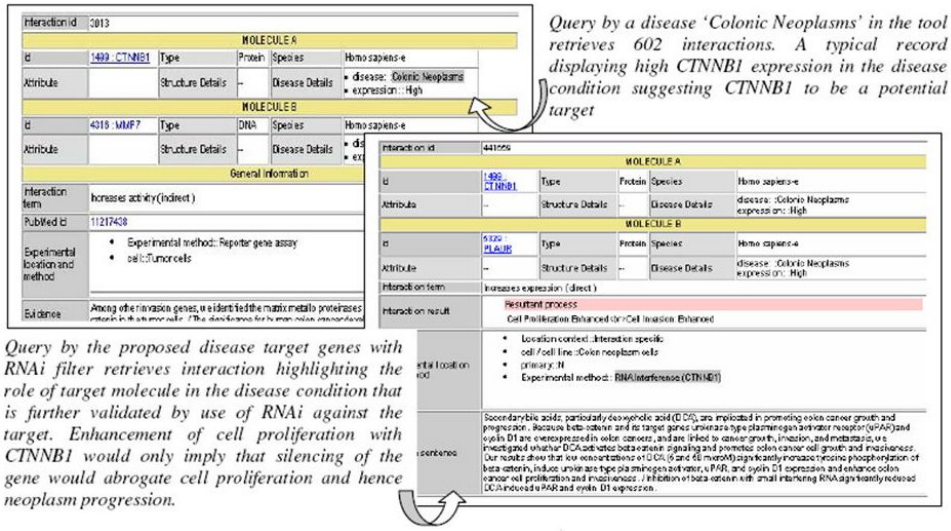


Figure 2: A typical result display.

Disease	Disease molecule association	Interaction	Interaction Result	Inference
Colonic Neoplasms	CTNNB1: High expression	CTNNB1 Increases expression CCND1	Cell Proliferation : Enhanced	CCND1 promotes cell cycle progression which leads to cell proliferation. As uncontrolled cell proliferation leads to neoplasm, CTNNB1 silencing by RNAi to decrease/inhibit CCND1 expression could be vital to stop the cell proliferation and hence the cancer [2].
Schizophrenia	AXIN1:potential target	AXIN1 Increases phosphorylation TP53	Apoptosis	Role of apoptosis in schizophrenia is well established [4]. TP53 being a tumor suppressor is involved in apoptosis [3]. Hence silencing of AXIN1 with RNAi will prevent TP53 function to induce apoptotic response in schizophrenia.
Osteoarthritis	ADAMTS4: involved	ADAMTS4 Cleave ACAN	Cartilage breakdown is induced	Aggrecan represents up to 10% of the dry weight of cartilage [5]. Aggrecan degradation is the root cause of Osteoporosis [6]. Prevention of Aggrecan cleavage by silencing ADAMTS4 could be beneficial in Osteoporosis.

Figure 3: Targets obtained in different diseases using the tool.

Cleavage of Mammalian Chromosomal DNA by Restriction Enzymes In Silico

Victor Tomilov,[1] Valery. Chernukhin,[1] Murat Abdurashitov,[1] Danila Gonchar,[1] and Sergei Degtyarev[1]

A theoretical method to simulate the digestion patterns of mammalian chromosomal DNA cleavage by restriction endonucleases was proposed. New software for long mammalian DNAs analysis using routine personal computers was developed. This computational technique includes short DNA sequences searching, DNA cleavage simulation, data treatment and verification.

Recently published primary structures of mammalian genomes (Rattus norvegicus, Mus musculus and Homo sapiens) were presented like databases and the analysis of short nucleotide sequences distribution in the corresponding genomes was performed. Computational DNA cleavage of genomes within the nucleotides sequences 5'-GGCC-3', 5'GATC-3', 5'-CC(A/T)GG-3' and 5'-CCGG-3', which are the recognition sites of well known restriction endonucleases (HaeIII, Kzo9I, Bst2UI and MspI respectively), was carried out and the diagrams of chromosomal DNA fragments distribution were obtained. Experiments on the chromosomal DNA digestion by corresponding restriction endonucleases were undertaken. The comparison of computational diagrams and results of chromosomal DNA cleavage was done and a high accordance of theoretical and experimental data was shown.

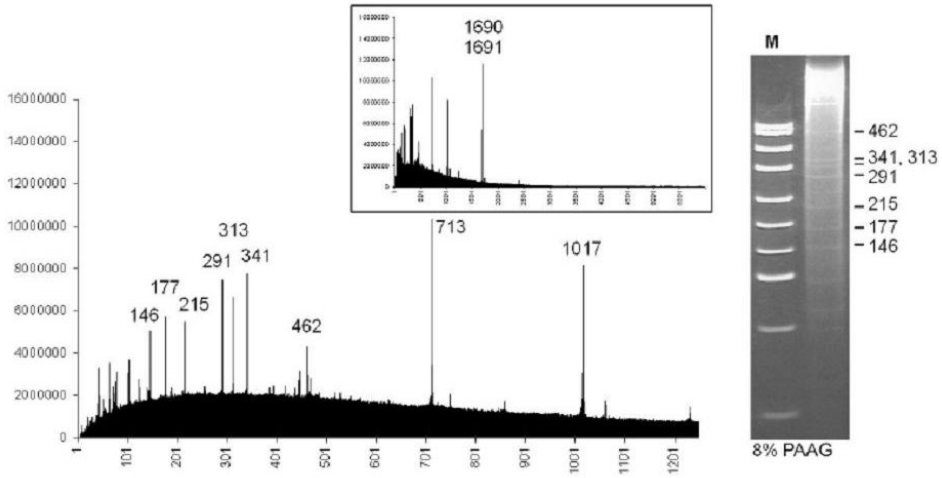


Figure 1: Distribution diagram example of total DNA fragments lengths (expressed in base pairs) depending on the fragment size for rat DNA cleavage at 5'-CC(A/T)GG-3' sequence. Shown on the right, there are the experimentally obtained patterns of respective DNA cleavage by restriction endonuclease Bst2UI (5'-CC(A/T)GG-3'recognition site).

[1]SibEnzyme Ltd., Novosibirsk, RUSSIAN FEDERATION

Comparative Analysis of Human Chromosomal DNA Digestion with Restriction Endonucleases In Vitro and In Silico

Murat Abdurashitov,[1] Victor Tomilov,[1] Valery Chernukhin,[1] Danila Gonchar,[1] and Sergei Degtyarev[1]

Theoretical analysis of human genomic DNA cleavage at 15 nucleotide sequences, which are the recognition sites of various restriction endonucleases, has been carried out. Distribution diagrams of calculated DNA fragments have been constructed based on earlier proposed method of mammalian genomes digestion in silico [1]. A similar study of human Alu- and LINE1-repeats digestion has been performed and corresponding diagrams of DNA fragments distribution have been plotted. Distribution diagrams of human genomic DNA digestion, which results in formation of low molecular weight DNA fragments, correspond to those for Alu-repeats; whereas the digestion, which results in formation of large molecular weight DNA fragments - are similar to those for LINE-repeats. All theoretical data have been compared to experimental patterns of human DNA hydrolysis with respective restriction endonucleases and a good correspondence for the most of DNA diagrams has been observed.

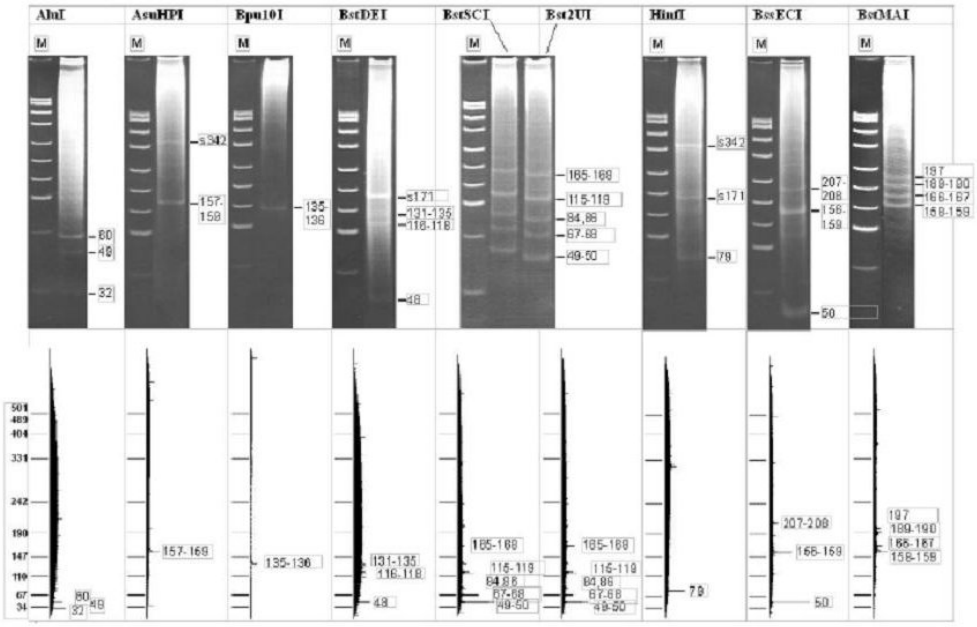


Figure 1: Example of comparison of electrophoregrams (8% PAAG) to calculated distribution diagrams of the total fragment lengths. Lengths of fragments with peak values, which can be determined at electrophoregrams are shown. "s" - fragments, which are probably result of satellite DNA cleavage. M - DNA fragment lengths marker pUC19/MspI. The lengths of fragments of this marker are shown at left in the bottom row.

References

[1] M.A. Abdurashitov, V.N. Tomilov, V.A. Chernukhin, D.A. Gonchar, S. Kh. Degtyarev. Cleavage of mammalian chromosomal DNA by restriction enzymes in silico. Online version at `http://science.sibenzyme.com/article14_article_27_1.phtml`.

[1]SibEnzyme Ltd., Novosibirsk, RUSSIAN FEDERATION

Granzyme B Cleavage Site Predictions based on Sequence, Physical Property and Structural Description of the Motif

Sebastian Maurer-Stroh,[1,2] Petra Van Damme,[3,4] Joost Van Durme,[3] Kim Plasman,[1,2] Evy Timmerman,[1,2] Pieter-Jan De Bock,[1,2] Marc Goethals,[1,2] Frederic Rousseau,[3] Joost Schymkowitz,[3] Joel Vandekerckhove,[1,2] and Kris Gevaert[1,2]

1 Introduction

Granzyme B (GrB) is an important apoptotic cytotoxic lymphocyte serine protease with, previously, only few known substrates. Recent elegant proteomics experiments resulted in the identification of a large number of substrate proteins cleaved by GrB in a cellular context [1].

2 Results

Based on this data we refined the characteristics of the GrB cleavage motif in terms of position-specific amino acid preferences, physical property constraints and pseudoenergies derived from structural modeling (Figure 1). In particular, we find limitations on the residue size at the immediate two positions flanking the cleavage site, as well as a region of preferred negatively charged or hydrophilic residues stretching from position P2 to P7.

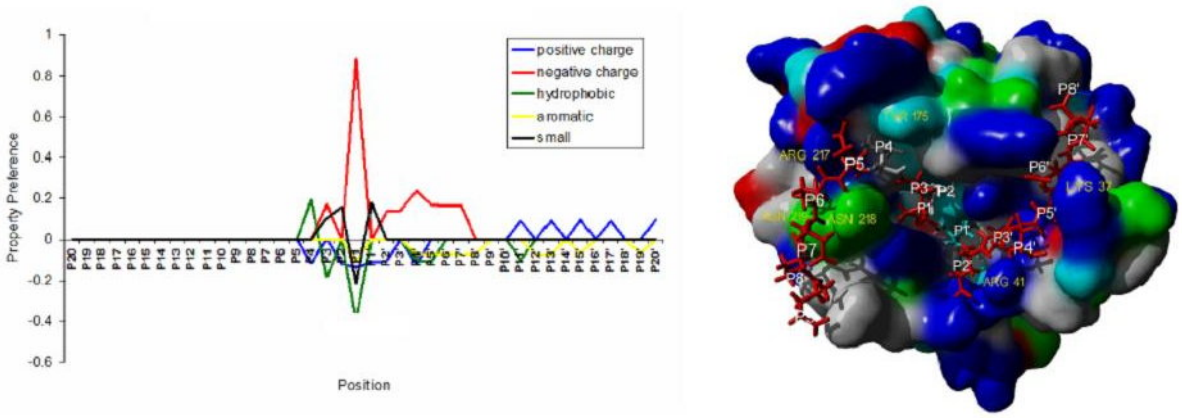


Figure 1: Physical property characteristics of motif (left). Structural model of extended substrate peptide in binding pocket of human GrB (right).

Compared to previous attempts on substrate prediction, we can now take advantage of a more complete motif description that includes an extension of the considered motif length from 4 (or 6) to at least 11 residues, hence we term our predictor GrB11. The wealth of new learning and training examples, coupled with an extension of our established prediction methodology [2], allowed for a significant jump in

[1]Switch Laboratory, VIB, VUB, Brussels, Belgium.
[2]Biomolecular Function Discovery Group, Bioinformatics Institute (BII), A*STAR, Singapore. Email: sebastianms@bii.astar.edu.sg
[3]Department of Medical Protein Research, VIB, Ghent, Belgium
[4]Department of Biochemistry, Ghent University, Ghent, Belgium

prediction performance of GrB cleavage motifs from the sequence under strict crossvalidation conditions (Figure 2). Especially, the rate of false positive predictions was found to be lowered, making our tool better suited for large database predictions.

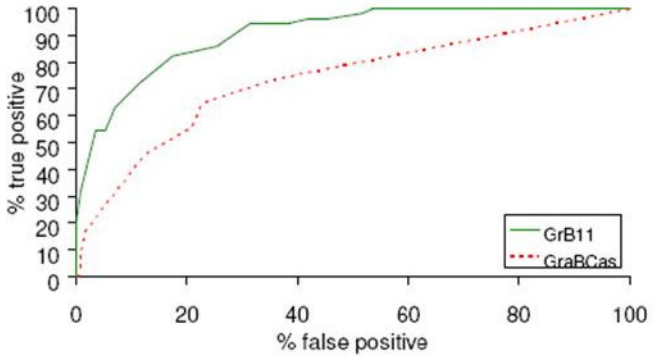


Figure 2: Crossvalidated ROC benchmark over new substrate data sets.

Moreover, we identified factors determining differences in substrate specificity between mouse and human GrB and developed a scoring function to predict these taxon specificities. We show that our scores correlate linearly with quantitative experimental data (Figure 3).

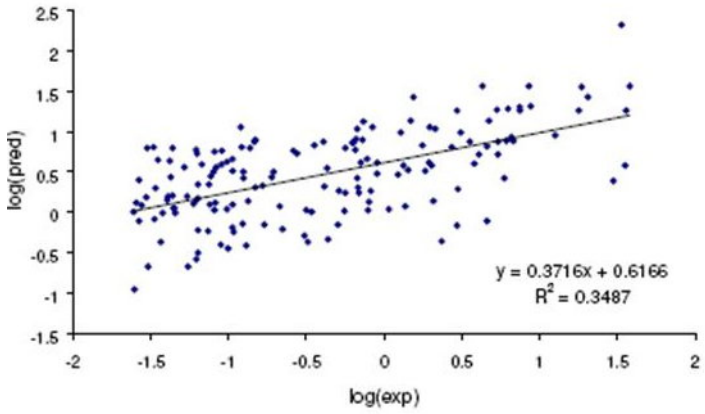


Figure 3: Correlation between predicted and experimental human to mouse substrate preference ratios.

3 Conclusions

We present GrB11, a new prediction tool for GrB substrates based on a vastly improved motif description utilizing sequence, physical property and structural information. GrB11 can even capture subtle differences in substrate specificity between human and mouse GrB.

References

[1] Van Damme P. et al. Differential substrate analysis by targeted proteomics reveals speciesspecific macromolecular substrate determinants of granzyme B. 2008, submitted.

[2] Maurer-Stroh S. et al. Refinement and prediction of protein prenylation motifs. *Genome Biol.*, 6(6):R55, 2005.

A Multi-Species Comparative Structural Bioinformatic Analysis of Inherited Mutations in α-D-Mannosidase

Javed Mohammed Khan,[1] Shoba Ranganathan[1,2]

1 Introduction

Lysosomal α-mannosidase is an enzyme that acts to degrade the N-linked oligosaccharides and hence plays an important role in mannose metabolism in humans and other mammalian species, especially livestock. Mutations in the MAN2B1 gene encoding lysosomal α-D-mannosidase cause improper coding, resulting in dysfunctional or non-functional protein and hence causing the disease α-mannosidosis. The phenotypic severity in this kind of inherited diseases is often found to be in correlation with the genotype. Mapping disease mutations to the structure of the protein can help in understanding the functional consequences of these mutations and thus indirectly, the finer aspects of the pathology and clinical manifestations of the disease in humans, cats, cows and guinea pigs.

2 Methods and Results

We performed a comprehensive homology modelling study and analysis of all the wild-type and mutated sequences of lysosomal α-mannosidase in four different species - human, cow, cat and guinea pig. Using the X-ray crystallographic structure of bovine lysosomal α-mannosidase (PDB ID 1O7D) [1] as the template to build the models of both wild type and mutated structures with all four disulfide linkages and bound ligands, we successfully established a satisfactory correlation between the severity of the genotype and that of the phenotype of the disease. Development of the detailed structural models required the use of several programs. CLUSTALX was used to align the target (wild-type) sequences with that of the bovine α-mannosidase template, MODELLER [3] was used to build the 3D structural models, while RASMOL was used for visualization and analysis. Quality checks and verification of the structural models were performed using the Biotech Validation Suite for Protein Structures web server, incorporating three major tools, PROCHECK, WHAT IF and PROVE. The WEBLOGO server was also used to find out the occurrence of the conserved residues and domains in all the species. The mutational, sequence, structure and literature data was obtained from OMIA [2], OMIM, Swiss-Prot, PDB and PubMed databases. In all, wild-type lysosomal α-mannosidase models for human, cat and guinea pig were generated, followed by mutant structures based on their respective wild-types: 11 for human, 2 for bovine and 1 for guinea pig [1, 4]. Several truncation mutations were also noted but structural models for these were not constructed. We have mapped all available mutations in the context of the enzyme active site in Fig. 1. Based on the analysis of structural models, we have correlated the position and functional consequence of the mutation to the observed phenotypic consequence (Table 1). All the truncation mutations and the mutations involving residues in and around the active site and also those destabilizing the fold led to severe genotypes and had lethal phenotypes as well. On the other hand, the mutations located distal to the active site were milder in both their genotypic and phenotypic expression.

3 Conclusion

This investigation highlights the importance of the proteins structure in its function and also forms the base for understanding the molecular reasons for the lethality or viability of the disease in different animal species. It proves that there is a significant correlation between the genotype and the phenotype

[1]Dept of Chemistry & Biomolecular Sciences, Macquarie University, Sydney, Australia. Email: jkhan@cbms.mq.edu.au
[2]Dept of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore. Email: shoba.ranganathan@mq.edu.au

of the disease. This study could play a vital role in drug designing and other therapeutic applications for inherited diseases.

Acknowledgment. JMK is grateful to Macquarie University for the award of an MQRES research scholarship.

References

[1] Heikinheimo, P., Helland, R., Leiros, H.-K.S., Leiros, I., Karlsen, S., Evjen, G., Ravelli, R., Schoehn, G., Ruigrok, R. and Tollersrud, O.-K. 2003. The Structure of Bovine Lysosomal α-Mannosidase Suggests a Novel Mechanism for Low-pH Activation. *Journal of Molecular Biology*, 327:631–644.

[2] Lenffer, J., Nicholas, F.W., Castle, K., Rao, A., Gregory, S., Poidinger, M., Mailman, M.D. and Ranganathan, S. 2006. OMIA (Online Mendelian Inheritance in Animals): An enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Research*, 34:D599–D601.

[3] Sali, A. and Blundell, T.L. 1993. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, 234:779–815.

[4] Sbaragli, M.. Bibi, L.. Pittis, M.G., Balducci, C., Heikinheimo, P., Ricci, R., Antuzzi, D., Parini, R., Spaccini, L., Bembi, B. and Beccari, T. 2005. Identification and characterization of five novel MAN2B1 mutations in Italian patients with alpha-mannosidosis. *Human Mutation*, 25:320.

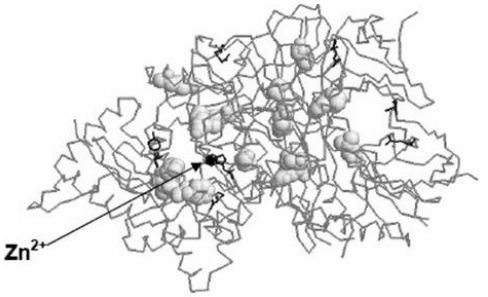


Figure 1: Structure of bovine lysosomal α-mannosidase showing the location of all the mutations studied. The residues in spheres represent the substitution mutations in different species (Table 1), while the residues represented as sticks highlight the frame-shift locations for all the truncation mutations. The catalytic zinc ion is shown as a black sphere.

S.No	Species	Mutated residues	Structural location	Effect on Structure	Phenotypic effect
1	Human	H72L	In AS	Destroys AS	Lethal
2	Human	H200N	Close to AS	Disrupts the fold	Harmful
3	Bovine	R220H	In AS	Destroys AS	Lethal
4	Guinea pig	R227W	Close to AS	Disrupts the fold	Harmful
5	Bovine	F320L	Close to AS	Disrupts the fold	Harmful
6	Human	T355P	Close to AS	Disrupts the fold	Harmful
7	Human	P356R	Close to AS	Disrupts the fold	Harmful
8	Human	E402K	Away from AS	Little or no effect	Viable
9	Human	S453Y	Away from AS	Little or no effect	Viable
10	Human	L518P	Close to AS	Disrupts the fold	Harmful
11	Human	W714R	Away from AS	Slight hindrance	Mild
12	Human	R750W	Away from AS	Slight hindrance	Mild
13	Human	G801D	Away from AS	Little or no effect	Viable
14	Human	L809P	Away from AS	Slight hindrance	Mild

Table 1: Structurally important mutated residues and their effect on folding and disease phenotype. AS: Active site.

Large-Scale Analysis and Screening of Chikungunya Virus T-cell Epitopes

Diane Simarmata,[1] Joo Chuan Tong,[2] Philippe Kourilsky,[3] Lisa F.P. Ng[4]

1 Introduction

Chikungunya fever is a viral disease transmitted by the Aedes mosquitoes. This disease shares similar symptoms with dengue such as high fever and rashes, but is distinguished by severe joint pains. Since 2006, massive outbreaks have been reported in the Indian Ocean Islands [1]. Globally, the disease has become a public health concern as the disease has also spread across the globe due to "imported" cases. The Chikungunya virus (CHIKV) belongs to the genus Alphavirus (Togaviridae family) and possesses a linear, positive-sense, single-stranded RNA genome of 11.8kb. It encodes two polyproteins - i) the non-structural polyprotein (nsP) for viral replication, and ii) the structural polyprotein consisting of 1 capsid protein (C), 2 major envelope surface glycoproteins (E1, E2), and 2 small structural proteins (E3, 6K) (Figure 1) [2].

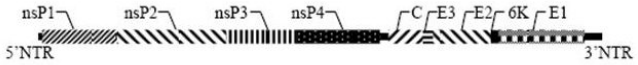


Figure 1: Genome structure of CHIKV.

At present, the role of T-cells in the pathogenesis of CHIKV remains unknown. In this study, we report the large-scale analysis and screening of CHIKV T-cell epitopes using an integrated ANNHMM predictive model. We examined whether 1) HLA-A2 class I alleles (A*0201, A*0202, A*0203, A*0204, A*0205, A*0206, A*0207, A*0209) show evidence of CHIKV peptide selection; 2) the extent of selection for peptides by the 8 different class I alleles; and 3) location of immunological hotspots (regions with high concentrations of T-cell epitopes) for each class I alleles.

2 Materials and Methods

2.1 Data

A total of 38 structural and 30 non-structural CHIKV sequences were extracted from Swiss-Prot [3]. From these, 73,247 nonameric peptide sequences (46,769 non-structural peptides, 26,478 structural peptides) were generated and used for the current analysis.

2.2 Algorithm

We used hidden Markov model (HMM) and artificial neural network (ANN) as the prediction engines. Each amino acid in a nonamer peptide is encoded as a binary string of length 20 with a unique position set to "1" and other positions set to "0". The outputs were binding scores ranging from 0 to 10, in increasing level of binding affinity (non-binding, N: 0.00-2.99; low binding affinity, L: 3.00-4.99; medium binding affinity, M: 5.00-6.99; high binding affinity, H: 7-10). For the HMM model, a first-order HMM is applied in which the current system state is determined only by the preceding state. For the ANN model, a 3-layer ANN was used. Peptides are predicted based on the consensus scores of both HMM and

[1]Singapore Immunology Network, A*STAR, Singapore. Email: diane_simarmata@immunol.a-star.edu.sg
[2]Institute for Infocomm Research, A*STAR, Singapore. Email: jctong@i2r.a-star.edu.sg
[3]Singapore Immunology Network, A*STAR, Singapore. Email: philippe_kourilsky@immunol.a-star.edu.sg
[4]Singapore Immunology Network, A*STAR, Singapore. Email: lisa_ng@immunol.a-star.edu.sg

ANN models. In this study, immunological hotspots are defined as regions within a sliding window of 30 amino acids that contain 4 or more predicted high-affinity binders. The ANN and HMM algorithms, training and testing were described in an earlier study [4].

3 Results

To examine whether HLA-A2 alleles show evidence of selection of CHIKV peptides, we screened 73,247 CHIKV nonameric peptide sequences (46,769 non-structural peptides, 26,478 structural peptides) for their ability to bind to 8 HLA-A2 (A*0201, A*0202, A*0203, A*0204, A*0205, A*0206, A*0207, A*0209) molecules. If an A2 allele is implicated in disease, we would expect a large proportion of the CHIKV peptides to be positively HLA associated. Of 26,438 structural CHIKV binding sequences, the number of A*0201, A*0202, A*0203, A*0204, A*0205, A*0206, A*0207, A*0209 predicted binding ligands are 45.3% (11,995/26,478), 68.4% (18,110/26,478), 66.56% (17,624/26,478), 53.7% (14,219/26,478), 72.52% (19,203/26,478), 63.56% (16,829/26,478), 49.29% (13,052/26,478), and 45.3% (11,995/26,478) respectively. A*0205 has the largest proportion of predicted binding ligands. 13 conserved A2-specific immunological (T-cell epitope) hotspots were predicted to exist (10 within the non-structural polyprotein; 3 within the structural polyprotein).

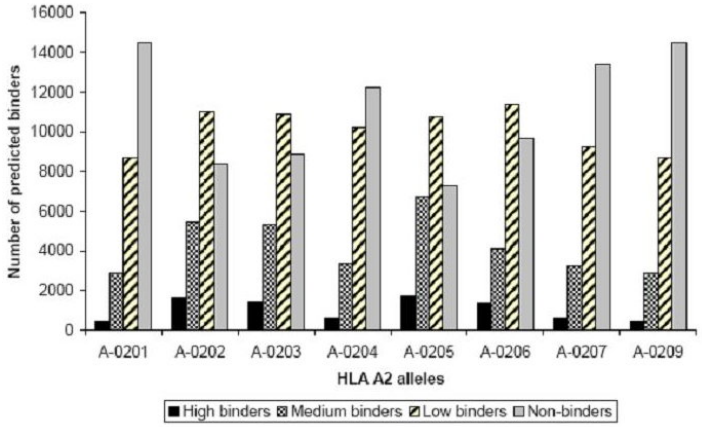


Figure 2: Number of predicted binding peptides for the 8 HLA-A2 alleles.

References

[1] Pialoux, G. et al. 2007. Chikungunya, an epidemic arbovirosis. *The Lancet Infectious Diseases*, 7:319–327.

[2] Powers, A.M. et al. 2000. Re-emergence of chikungunya and o'nyong-nyong viruses: Evidence for distinct geographical lineages and distant evolutionary relationships. *Journal of General Virology*, 81:471–479.

[3] Boeckmann, B., et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31:365–370.

[4] Srinivasan, K.N., et al. 2004. Prediction of class I T-cell epitopes: evidence of presence of immunological hot spots inside antigens. *Bioinformatics*, 20:i297–i302.

Statistical Analysis of KMSKS Motif in Aminoacyl-tRNA Synthetase by Building a Library of Random Sequences

Shunsuke Kamijo,[1] Akihiko Fujii, Kenji Onodera, Kenichi Wakabayshi, Takatsugu Kobayashi, and Kensaku Sakamoto[2]

1 Introduction

KMSKS loop is a symbolic name of well known motif which is highly related to ATP reactions in aminoacyl-tRNA synthetase, and the loop is assumed to be responsible for acquiring energy in aminoacylation process [3, 4]. The amino acid sequences of KMSKS loop are conserved with small mutations among aminoacyl-tRNA synthetase of different species, and some analyses on the mechanism of the loop have been performed [3, 2, 1]. However, concrete rules for the loop to keep the activation have not been revealed yet.

Today's bioinformatics approaches generally employ statistical analyses of the databases such as PDB(Protein Data Bank), and those approaches have been proved to be quite successful in finding motifs and other conserved sequences important for the activation of proteins. However, most of genome information in the public databases was acquired from several natural proteins, and such the amino acid sequences can be regarded as one of the optimal sequences. Therefore it would be quite significant to build library of mutants having other sequences for the motif than that of natural proteins as many as possible, and to analyze correlations between the motif sequences and the degrees of activities of the mutants.

2 Building a Library of KMSKS Mutants

In order to find such the rules for KMSKS loop, the loop was replaced by the random sequences made of five amino acids in the tyrosyl-tRNA synthetase(TyrRS), and a library of about a hundred of mutants of TyrRS was build. Here the corresponding sequence of KMSKS loop in TyrRS of Methanococcus jannashii is 'KMSSS', and 'EGKMSSSKG' is a sequence including neighboring amino acids of 'KMSSS'. In this work, 'EGKMSSSKG' was replaced by random sequences.

Activities of the mutants were measured by the Amber suppression method. The genome for the chroramphenicol resistant protein was coded in the plasmid, and the plasmids are transformed into competent cells. However, for this method, a condon for tyrosine in the genome is replaced by Amber codon., and the genome of tRNA designed to install tyrosine into the position indicated by Amber codon 'UGA' is also cloned into the plasmid. If a mutant of TyrRS has an activity to combine tyrosine at the place indicated by Amber codon, the chroramphenicol resistant protein will be synthesized, and the competent cell will obtain resistance to chroramphenicol. Our library is obtained from colonies survived on culture media contained chloramphenicol of $300mu$g/ml.

By this method, a lot of substitutable sequences for KMSKS can be obtained, and those complemental data should be quite suggestive for informatics analyses. In this work, we performed a statistical analysis of possible sequences of KMKSK loop.

3 Statistical Analysis of Mutational Expressions of KMSKS Loop

Table 1 shows numbers of colonies survived on the culture media containing 300g/ml chloramphenicol. From the statistics, some important rules were suggested as follows. (1) 'K' at position 1 was not replaced

[1]The University of Tokyo, Institute of Industrial Science. Email: kamijo@iis.u-tokyo.ac.jp
[2]RIKEN, Genomic Science Center. Email: sakamoto@gsc.riken.jp

by other amino acid residues. (2) Position 2 has a partial restriction for the mutations to keep activities. (3) 71% of expressions at position 3, 4, and 5 were occupied by 'S', 'G', and 'A'. (4) Residues at positions of P2, P1, B1, and B2 have only a few restrictions for the mutations to keep activities. (5) 'E' or 'D' did not appear at any positions.

From the above statistics, some rules were concluded as follows. (1) Positively charged side chains are important to bind ATP which is charged negatively because of phosphate groups. (2) Negatively charged side chains are not allowed because they should repel the ATP with negatively charged phosphate groups. (3) 'S', 'G', and 'A' are preferable in order to keep the loop flexible due to their small side chains with poor reactivity.

Above rules are important for binding ATP by the loop. For the future work, we will continue to study to reveal the dynamics how ATP is hydrolyzed by the KMSKS loops.

References

[1] Austin, J. and First, E.A. 2002. Comparison of the catalytic roles played by the KMSKS motif in the human and Bacillus stearothermophilus tyrosyl-tRNA synthetases. *Journal of Biological Chemistry*, 277:28394–28399.

[2] First, E.A. and Fersht, A.R. 1993. Involvement of threonine 234 in catalysis of tyrosyl adenylate formation by tyrosyl-tRNA synthetase. *Biochemistry*, 32:13644–13650.

[3] Hountondji, C., Dessen, P. and Blanquet, S. 1986. Sequence similarities among the family of aminoacyl-tRNA synthetases. *Biochimie*, 68:1071–1078.

[4] Kobayashi, T., Nureki, O., Ishitani, R., Yaremchuk, A., Tukalo, M., Cusack, S., Sakamoto, K. and Yokoyama, S. 2003. Structural basis for orthogonal tRNA specificities of tyrosyl-tRNA synthetases for genetic code expansion. *Nature Structural Biology*, 10:425–432.

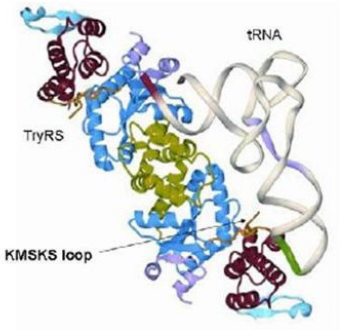


Figure 1: KMSKS Loop in the structure of TyrRS.

Position	Original Expression	Expressions from survived colonies
P2	E	(4)Y, (3)G, (2)LTVR, (1)IHFAQMWCSP
P1	G	(4)G, A, (3)T, V, (2)W, I, R, (1)LNCKY
1	K	(41)K
2	M	(7)M, (3)A, C, (2)Q, (1)S
3	S	(22)S, (20)G, (11)A, (2)C, (1)T
4	S	(22)S, (8)A, (6)T, (3)CV, (1)KN
5	S	(23)S, (12)G, (6)RA, (5)C, (4)T, (3)FVML, (2)QYH, (1)W K I
B1	K	(7)G, (6)LR, (3)AS, (2)WN, (1)HVFY
B2	G	(7)G, (4)L, (3)RAV, (2)SWYCN, (1)KDN

Table 1: Numbers of Expressions at each position around the KMSKS loop. In the above table, '(20)G' in the row labeled as 'position 3' means that 'G' was found from twenty colonies at position 3 where the expression was originally 'S'.

Whole Genome Assembly from 454 Sequencing Output[1]

Jacek Blazewicz,[3,4,2] Marcin Bryja,[3] Marek Figlerowicz,[4] Piotr Gawron,[3]
Marta Kasprzak,[3,4] Darren Platt,[5] Jakub Przybytek,[3]
Aleksandra Swiercz,[3,4] Lukasz Szajkowski[5]

1 Introduction

The DNA sequence assembly, one of the most important problems of computational biology, is well known for its high complexity, due to huge amount of erroneous and incomplete data. Many teams worldwide put their efforts to provide heuristics producing satisfying semi-optimal outcomes [3, 4, 6, 8]. The errors present in the data come from the previous stage in the process of recognizing genetic information of organisms, namely the DNA sequencing. Recently a new biochemical method of DNA sequencing, 454 sequencing owned by 454 Life Sciences Corporation, has been introduced [5]. It gives highly reliable output of low cost and in short time. 454 sequencing is based on the pyrosequencing protocol [7]. For assembly purposes this method is much better than the others from the point of view of sequence reliability. On the other hand, its sequences are usually of length 100–200 nucleotides while fragments produced by other sequencing methods are of length of a few hundreds of nucleotides.

The specificity of the data from the 454 sequencing impacts on an assembly algorithm used in the computational stage. A new assembly algorithm has been proposed which deals well with these data and outperforms other assembly algorithms known from the literature. The algorithm is a heuristic based on a graph model, the graph being built on the set of input sequences. The computational tests were performed on the data coming from real biochemical experiment, done in Joint Genome Institute in order to sequence the whole genome of bacteria *Prochlorococcus marinus*.

2 Method and Discussion

In the proposed algorithm—SR-ASM (Short Reads ASseMbly)—three parts can be distinguished. The first part computes feasible overlaps for all input sequences. It requires as the parameters, values of the minimum overlap between two sequences and of the error bound, the latter being the percentage of the mismatches allowed in the overlap of two sequences. The comparison is done also for the reverse complementary sequences to the input ones, due to the assumption that the fragments come from both strands of a DNA helix. In the second phase, a graph is constructed with the fragments as the vertices, and two vertices are connected by an arc if there is a feasible overlap between the two fragments. Next, a path is searched for, which passes through one of the vertices from every pair: either through the straightforward fragment or its reverse complementary counterpart. Usually it is not possible to find a single path in the graph and as the solution several paths corresponding to contigs are returned. At the end, in the third part, a consensus sequence (sequences) is determined on the basis of the alignment.

The proposed new assembly algorithm has been tested on raw data coming from real experiment with the 454 sequencer, done in Lawrence Livermore National Laboratory operating within Joint Genome Institute. The data covered the whole genome of bacteria *Prochlorococcus marinus*, of length 1.84 Mbp. The output of the sequencer contained above 300000 sequences of length about 100 nucleotides. Together with the sequences their rates of confidence were provided. For the computational experiment, also smaller sets of input sequences were prepared in order to compare the behavior of distinct methods designed for the genome assembly problem. The model sequence was known and published [2]. However, it was used only after the computations, for the comparison of the similarity with the obtained contigs.

Several publicly available algorithms were tested, among them PHRAP (http://www.phrap.org/), CAP3 (http://genome.cs.mtu.edu/), TIGR (http://www.tigr.org/) and our previous assembly program

[1]The research has been partially supported by a Polish Government grant.
[2]Corresponding author: jblazewicz@cs.put.poznan.pl
[3]Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland.
[4]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland.
[5]Lawrence Livermore National Laboratory, Joint Genome Institute, 7000 East Avenue, Livermore, CA 94550, USA

ASM [1]. The tests of the algorithms were carried out on SUN Fire 6800 in Poznan Supercomputing and Networking Center. The results of tests for algorithms SR-ASM, and PHRAP (produced the best solutions except from our algorithm) are presented in Table 1. The criteria used to compare the outcomes of the methods are: the number of contigs obtained by the methods, the quality of the largest contigs (i.e. the similarity of a contig to a fragment of the genome), the coverage (the percentage of the total length which was covered by the contig), and the time of computations. For the cases where the algorithms resulted in more than one contig, the coverage and quality values are presented for three longest contigs. In the table, "in-s" means the number of input sequences in instances, "in-c" means the number of model contigs, "qual" stands for the quality, "cov" for the coverage, and "cntg" for the number of contigs obtained by the assemblers.

Among the methods—PHRAP, ASM, CAP3, TIGR, and SR-ASM—our algorithm appeared to be the best both in the number of produced contigs and in the lengths of the contigs. The computational time of algorithm SR-ASM is rather long, but the time is not crucial for assembly algorithms. SR-ASM solved the whole genome in 80 hours, what is quite acceptable with regard to the time of obtaining these data in biochemical experiments. It should be noticed that we have also a parallel implementation of the same algorithm, which significantly reduces the time.

References

[1] Blazewicz J., Figlerowicz M., Formanowicz P., Kasprzak M., Nowierski B., Styszynski R., Szajkowski L., Widera P., and Wiktorczyk M. Assembling the SARS-CoV genome—new method based on graph theoretical approach. *Acta Biochimica Polonica*, 51:983–993, 2004.

[2] Chen F., Alessi J., Kirton E., Singan V., and Richardson P. Comparison of 454 sequencing platform with traditional Sanger sequencing: A case study with de novo sequencing of Prochlorococcus marinus NATL2A genome. In: *Plant and Animal Genomes Conference*, 2006.

[3] Idury R. andWaterman M. A new algorithm for DNA sequence assembly. *Journal of Computational Biology*, 2:291–306, 1995.

[4] Kececioglu J.D. and Myers E.W. Combinatorial algorithms for DNA sequence assembly. *Algorithmica*, 13o:7–51, 1995.

[5] Margulies M., Egholm M., Altman W.E., Attiya S., Bader J.S., et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380, 2005.

[6] Pevzner P.A., Tang H., and Waterman M.S. A new approach to fragment assembly in DNA sequencing. In: *Proc. of 5th Annual International Conference on Computational Molecular Biology* (RECOMB'01), Montreal, pages 256–267, 2001.

[7] Ronaghi M., Karamohamed S., Pettersson B., Uhlen M., and Nyren P. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, 242:84–89, 1996.

[8] Sundquist A., Ronaghi M., Tang H., Pevzner P., and Batzoglou S. Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS ONE*, 2:e484, 2007.

		SR-ASM				PHRAP		
in-s	in-c	cntg	qual [%]	cov [%]	time [s]	cntg	qual [%]	cov [%]	time [s]
1000	1	1	93.49	100.00	5.0	4	96.93	90.43	3.5
							99.32	9.66	
							88.85	3.95	
5000	1	1	94.83	100.00	31.0	7	97.78	99.27	19.2
							88.85	0.80	
							98.10	0.63	
10000	1	1	94.89	100.00	113.0	10	99.60	47.98	37.2
							99.51	34.35	
							99.52	10.38	
50000	21	31	95.18	10.17	5926.0	63	99.59	11.62	218.6
			92.56	10.10			96.49	7.40	
			97.27	9.13			97.95	6.94	
100000	42	76	75.95	10.42	22987.0	135	73.17	6.74	487.7
			95.40	9.60			98.69	6.48	
			95.14	6.94			99.60	5.82	
150000	79	132	79.48	7.70	57241.0	225	98.70	4.32	782.9
			75.96	6.96			79.58	3.69	
			95.40	6.41			99.65	3.32	
300000	150	171	95.34	5.66	289513.0	5460	94.56	4.74	2067.9
			94.91	4.38			81.66	1.89	
			95.07	3.19			98.56	1.88	

Table 1: Results of the computational experiment for algorithms SR-ASM and PHRAP.

ConView: An Easy and Fast Visualization Tool for Contig Assembly

Hongseok Tae,[1] Kiejung Park[2]

1 Introduction

Contig assembly is an important task in the genome sequencing project. To support the high throughput of the genome project, several contig assembly programs, such as Phrap [5] and ARACHNE [1], have been developed.

The finishing process, which closes gaps and improves the quality of the data, is the most time consuming step in genome assembly because of base-calling errors, contaminations and repeated sequences. To aid the contig assembly, visualization of a contig is essential. Finishing programs, such as Consed [4] and Gap4 [2], which were developed to simplify contig assembly, provide user interfaces for editing and annotating DNA sequences and navigating chromatograms of reads under the graphical environment. Most visualization programs for the contig assembly run on the LINUX platform or have a complicated user interface, requiring much effort to learn how to use it. Moreover, these programs require considerable time and computer memory to open contig data.

We have developed an easy and fast contig viewer, ConView, to aid the contig assembly. It runs on Windows PC and can read data in other platforms. To overcome the time and memory problems, a few efficient techniques were implemented.

2 Features and Results

ConView reads an ACE file, which is an output file of Phrap, and lists contigs with horizontal bars proportional to their relative lengths; Fig. 1(A). The ACE file contains overall contig information, including the reads composition of the contigs. A detailed information window for each contig is displayed by clicking the contig bar.

The Fragment representation mode shows the reads composition of the contig. In Figure 1(B), the black line on the top represents the contig and the other bars represent reads. As a contig is assembled using the high quality regions of reads, the regions are colored green on the read bars, whereas low quality regions are colored red. The sequence representation shows the DNA sequences of the contig and reads; Fig. 1(C). The ivory colored bases are the consensus sequence for the contig. Green and yellow represent high and low scores in the base-calling results, respectively, whereas red represents bases that are inconsistent with the consensus. The chromatogram representation reads the ABI chromatogram files, which are generated by ABI DNA Analyzer, containing the fluorescence trace data of the four DNA bases and shows their graphs on the reads positions; Fig. 1(D). The chromatograms are the raw data for the base-calling programs, such as Phred [3], that identify the DNA sequences of reads. Because most contig viewer programs operate on the LINUX platform, they are not easy for most genome researchers to learn. ConView, however, runs on Windows PC and has a user-friendly interface.

Since an ACE file contains all contigs for a genome, it requires a considerable amount of time to read contig data. Moreover, a huge contig consists of numerous reads, making it time consuming to arrange reads according to their positions. We have solved these problems by making a hash file, which contains the positions of contig data in the ACE file and the arrangement information of reads on the contigs. Another issue in contig visualization is the overhead of reading chromatogram data while navigating and sliding a contig. We have tried to reduce the waiting time and computer memories required to read chromatogram files. ConView allocates memory for the chromatogram files of reads around the currently displayed screen. While the window is sliding, it calculates the read composition displayed on the screen. If a read is no longer included on the current screen, the memory allocated for its chromatogram is released. When it runs files on a remote computer, it only transfers ABI files of reads currently displayed.

[1]Department of Computer Engineering, Chungnam National University, South Korea. Information Technology Institute, SmallSoft Co., Ltd., Daejeon, South Korea. Email: mbio94@naver.com

[2]Information Technology Institute, SmallSoft Co., Ltd., Daejeon, South Korea. Email: kjpark@smallsoft.co.kr

ConView allows researchers to navigate chromatograms at high speed without loading all chromatogram files of the contig in the memory.

3 Discussion

ConView contains several new features not contained in other programs. In developing this program, we have focused on a user-friendly interface and fast execution, features that can allow the researcher to work in a more comfortable environment. ConView has been developed as a visualization program that can be easily integrated into our genome annotation system. Although the current version of ConView reads only ACE files, the next version, now in development, will read various assembly files.

References

[1] Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., et al., 2002. ARACHNE: A whole-genome shotgun assembler. *Genome Research*, vol. 12, no. 1, pp. 177–189.

[2] Dear, S., and Staden, R. 1991. A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Research*, vol. 19, no. 14, pp. 3907–3911.

[3] Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, vol. 8, no. 3, pp. 175–185.

[4] Gordon, D., Abajian, C., and Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Research*, vol. 8, no.3, pp. 195–202.

[5] Green, P., 1995. Phrap and Cross_Match at `http://www.phrap.org`.

Figure 1: The interfaces of ConView. (A) The main page after an ACE file is imported. Horizontal bars represent relative lengths of contigs. (B) The fragment representation mode, which displays the reads composition of a contig. (C) The sequence representation mode, which displays the DNA sequences of the contig and reads. (D) The chromatogram representation mode, which displays fluorescence graphs of the reads.

Deciphering Functional Linkages between Mycobacterium Tuberculosis H37Rv Proteins via Gene Ontology Similarity Scores

Bhakti Bhagwat, Santosh Atanur, Sunitha Manjari, and Rajendra Joshi[1]

1 Introduction

The functional genomic approaches lead researchers to focus into the dynamic aspects of biological processes of living organisms such as translation, protein-protein interactions and functional linkages. The present study is an effort to decipher functional linkages between the proteins of *Mycobacterium tuberculosis* H37Rv (MTH) using Gene Ontology terms. The obtained maps of protein-protein functional linkages provide a valuable insight into the role of genes involved in several biochemical pathways.The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism. Gene Ontology Structure is organized in Directed Acyclic Graph (DAG) structure. In the DAG, the root node is "Gene Ontology" which is followed by three nodes Cellular Component (CC), Biological Processes (BP), Molecular Function (MF). These DAG structures of a pair of GO terms were used to measure the similarity between them in terms of relative specificity similarity(RSS). The RSS of a gene pair is considered as the measure for clustering the genes of MTH. There are 2554 genes with GO annotations in MTH of which 2014 genes have been classified under the GO term "Biological Process". These 2014 genes were used for further computations of RSS leading to the generation of a data set constituting ∼20 million gene pairs. This data set has been statistically validated with data from KEGG and METACYC databases. The Kolmogorov Smirnov (KS) Test was carried out to check how much a test dataset differs from training set. The KS test results show that at the D value of 0.33, the two data sets differ from each other significantly at RSS value of 0.46. Further, Z score analysis was carried out to determine the statistical significance of RSS values with respect to the assignments of protein pairs into different categories. The data set was divided into 10 data bins (ranging from 0.1 to 1.0) to carry out the Z-score test. The confidence of the strength of the relationship between protein pairs measured by RSS method was found to be higher in the GO bins > 0.8.

Based on statistical analysis, the genes were further clustered according to their RSS values into different protein networks using two clustering approaches viz., Recursive and Gene centric. Recursive clustering delineated the entire data set in to only 25 clusters as compared to gene centric clustering where 334 clusters were detected. This number is close to that observed in KEGG and METACYC for MTH (∼200). Further manual analysis of gene centric clusters refined them to 139 clusters.

Analysis of the manually refined gene centric clusters helped in the elucidation of novel pathways as well as assignment of new genes to existing pathways. For instance, in Fatty Acid Biosynthesis pathway, there are 42 genes in the cluster of which 19 newly predicted genes were associated with Mycolic acid biosynthesis, which clearly depicts the linkage between Fatty acid Biosynthesis and Polyketide Synthase (PKS) of Mycolic acid Biosynthesis (as shown in Figure 1). This has significance as polyketide synthase has definite role in pathogenesis of MTH. In carotenoid biosynthesis cluster, 4 new genes (viz., Rv3829c, Rv0897c, Rv1432 and Rv2997), which are not previously reported in KEGG and Metacyc, were identified. It is interesting to know that the orthologues of these genes have been reported to be part of carotenoid biosynthesis in *Mycobacterium marinium*.

Detection of functional linkages using semantic similarity measures provides a different perspective to understand biological phenomena. However, caution needs to be exerted while interpreting the relationship between the genes because the accuracy of the method is as good or as bad as the GO annotations provided by the existing tools. Prediction of functional modules using GO terms if applied in conjunction with other approaches like phylogenetic profiling and gene neighborhood analysis provides more comprehensive knowledge for detection of protein networks. Such a methodology has not only been used to validate the gene expression data but also helps in the identification of false positives in protein interaction studies.

[1]Bioinformatics Team, Scientific Engineering and Computing Group, Centre for Development of Advanced Computing, Pune University Campus, Ganesh Khind, Pune 411 007, India. Email: rajendra@cdac.in

2 Software and Files

InterProScan and Blast2GO were used together to annotate the genes from MTB with respect to GO terms. For further statistical analysis and visualization of clusters, the in-house JAVA based software tool is used. For back-end storage, MySQL Data base management system is used.

References

[1] Damien Portevin, Ce'lia de Sousa-D'Auria, Christine Houssin, Christine Grimaldi, Mohamed Chami, Mamadou Daffe', and Christophe Guilhot. 2004. A polyketide synthase catalyzes the last condensation step of mycolic acid biosynthesis in mycobacteria and related organisms. *PNAS*, vol. 101, pp 314–319.:w

[2] Hongwei Wu, Zhengchang Su, Fenglou Mao, Victor Olman, and Ying Xu. 2005. Prediction of functional Modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Research*, Vol. 33, No.9 pp 2822–2837.

[3] Lalita Ramakrishnan, Hien T. Tran, Nancy A. Federspiel, and Stanley Falkow. 1997. A crtB homolog essential for photochromogenicity in Mycobacterium marinum: Isolation, characterization, and gene disruption via Homologous Recombination. *Journal of Bacteriology*, Vol. 179, No.18.

[4] Xiaomei Wu, Lei Zhu, Jie Guo, Da-Yong Zhang and Kui Lin. 2006. Prediction of yeast protein-protein interaction network: Insights from the Gene Ontology and annotation. *Nucleic Acids Research*, Vol. 34, No. 9, pp 2137–2150.

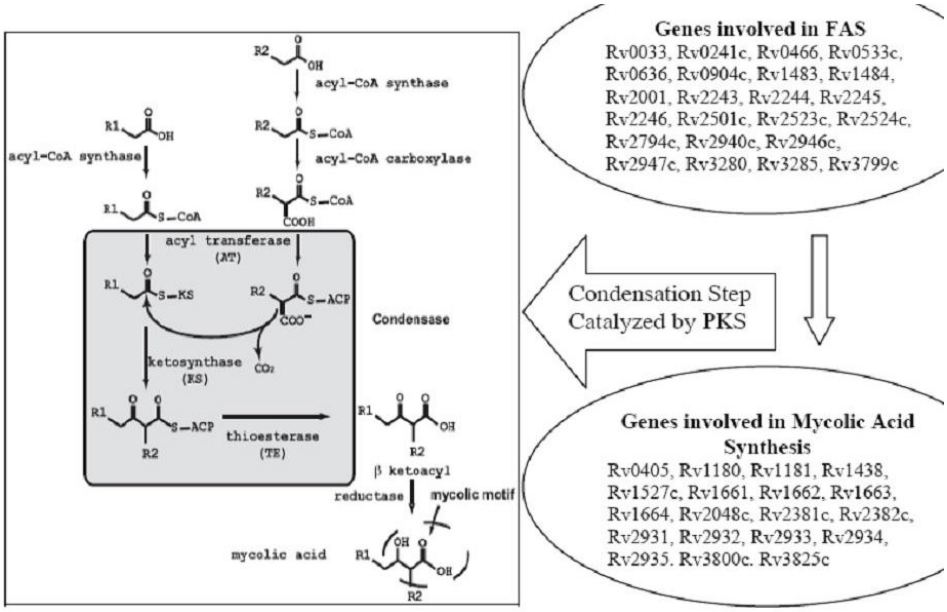


Figure 1: Functional interlinking between Fatty Acid Biosynthesis and Mycolic Acid Biosynthesis. Source: [1].

Planes in 3-Dimensional Metabolite Triplet Data: A Robust, Bayesian Approach

Bruno Schwenk,[1] Joachim Selbig,[2] Matthias Holschneider[3]

1 Data-Points and Planes

The correlation coefficient is a basic tools to asses a linear relationship between two metabolites in metabolic data analysis. It can be interpreted in terms of the quality of fit of that can be obtained by comparing the sample of data points in 2-dimensional space with all possible lines therein. We try to generalize this by investigating planes in a 3-dimensional data-space. This allows to detect triplets of metabolites, which are linearly covarying in the sense that one of them is a linear function of the other two.

2 Methods

The strategy we employ is to compare the set of all possible planes to the data points by computing a *Bayesian probability density function* (PDF) over a parameterization of the planes [2]. The comparison is based on a Gaussian *error model* which allows to obtain the expression for the PDF analytically as well as to include correlated errors into the analysis.

To ensure *robustness* against outlying values we calculate the PDF by a formula that includes the possibility of outliers and in consequence is intrinsically robust [3]:

$$PDF = \sum_{i=1}^{m} \ln \left[\exp\left(-\frac{1}{2} \frac{(\vec{x_i}\vec{n} - \beta)^2}{\vec{n}^T \sum \vec{n}} \right) + c \right]$$

A single plane is defined as the set of all points $\vec{x}$ that, for a given vector $\vec{n}$ of unit length and a real, positive number $\beta \geq 0$, fulfills the relationship:

$$(\vec{n}\vec{x} - \beta)^2 = 0$$

In consequence the set of all planes is the Cartesian product of the unit sphere resulting from $\vec{n}$ with the positive part of the real axis. An important point in applying Bayesian principles to planes is the calculation of the *invariant measure* of this set [1], which was analytically feasible in our case, and the definition of a dense, quasi uniform grid of discretization-points.

Fig. 1 visualizes the sphere part of the PDF, for a given value of β, that results from the generic case of a single data point in 3-dimensional space.

As part of a data analytical software tool we implemented and investigated various test statistics in form of functionals of the Bayesian PDF, such as differential entropy, a polynomial measure of concentration and curvature based statistics, that allow to detect if certain, plane related hypotheses are valid for the given data.

3 Application to Metabolic Data

To demonstrate the practicability of the method we analyzed a set of *triplets of metabolite intensities*, selected from data already published in [4], by comparing the values of different test statistics with distinct, plane related hypotheses (namely: data lie on a plane, a line, a single point). The errors that entered the calculations were highly correlated and based on the covariance matrix of the technical replicates of the metabolite measurements. We found that in most of the triplets under investigation data were clearly located on a straight line, in a space that's metric was given by the error model.

[1]Institut fur Mathematik, Universitat Potsdam, Germany. Email: schwenk@math.uni-potsdam.de
[2]Max-Planck-Institut fur Pflanzenphysiologie, Golm, Germany. Email: selbig@mpimp-golm.de
[3]Institut fur Mathematik, Universitat Potsdam, Germany. Email: matthias.holschneider@googlemail.com

References

[1] E.T. Jaynes. The Well-Posed Problem. *Foundations of Physics*, Vol.3, No.4, 1973.

[2] E.T. Jaynes, G.L. Bretthorst. Probability Theory. *The Logic of Science: Theory and Elementary Applications, Vol. 1*. Cambridge University Press, 2003.

[3] Frank Kose, Jan Budzies, Matthias Holschneider, Oliver Fiehn. Robust detection and verification of linear relationships to generate metabolic networks using estimates of technical errors. *BMC Bioinformatics*, 8:162, 2007.

[4] M. Scholz, S. Gatzek, A. Sterling, O. Fiehn, J. Selbig. Metabolite fingerprinting: Detecting biological features by independent component analysis. *Bioinformatics*, 20(15):2447–2454, 2004.

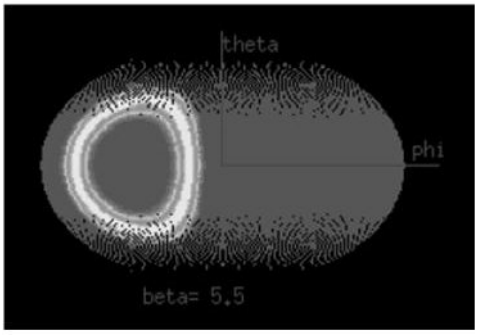


Figure 1: Spherical part of the Bayesian probability-density-function (PDF), resulting from a given data-point $\vec{x} = (0, 10, 0)$ for a value of $\beta = 5.5$ (beta slice). The error model is isotropic with a standard deviation of $\sigma = 1$.

Deciphering Transmembrane Helix Recognition by the ER Translocon

Tara Hessa,[1*] Nadja M. Meindl-Beinker,[1*] Andreas Bernsel,[2*] Hyun Kim,[1] Yoko Sato,[1] Mirjam Lerch-Bader,[1] IngMarie Nilsson,[1] Stephen H. White[3] and Gunnar von Heijne[1,2]

1 Introduction

Most integral membrane proteins are composed of tightly packed transmembrane (TM) α-helices, which are recognized and inserted into the membrane co-translationally by complex molecular machines called translocons. The physicochemical characteristics of the membrane vary markedly over short distances, which is reflected in the distribution of different amino acids in the membrane-embedded parts of integral membrane proteins. But what is exactly the "molecular code" that allows a translocon to recognize TM helices in newly synthesized polypeptide chains?

Here, we have analysed a large number of systematically designed transmembrane α-helices for their insertion efficiency into dog pancreas rough microsomes in an *in vitro* translation system, Fig. 1.

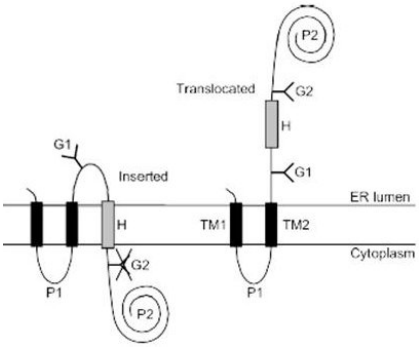


Figure 1: The model protein has two TM helices (TM1 and TM2) and a large luminal domain (P2). Systematically designed TM-helices (H; grey) are engineered into the P2 domain with two flanking glycosylation acceptor sites (G1, G2). Constructs for which the H-segment is integrated into the ER membrane as a TM helix are glycosylated only on the G1 site (left), whereas those for which the H-segment is translocated across the membrane are glycosylated on both the G1 and G2 sites (right).

2 A Quantitative Model for Transmembrane Helix Recognition

From the insertion efficiency data, a quantitative model for TM helix recognition by the endoplasmic reticulum translocon was developed, in which amino acid contributions to the overall free energy of insertion were assumed to be dependent on position along the membrane normal. Profiles describing the contribution from each amino acid as a function of sequence position were represented by gaussian functions (Fig. 2), the parameters of which were optimized to minimize the difference between experimentally measured values and calculated values according to an additive model. The profiles resulting from the optimization (Fig. 2; black curves) are similar to statistical free-energy curves (Fig. 2; grey curves), derived from the distribution of the different amino acids along the membrane normal in high-resolution membrane protein 3D structures; these profiles presumably reflect mainly interaction free energies between amino-acid side chains and the lipid bilayer. The effects from TM segment length and flanking amino acids were also incorporated into the model.

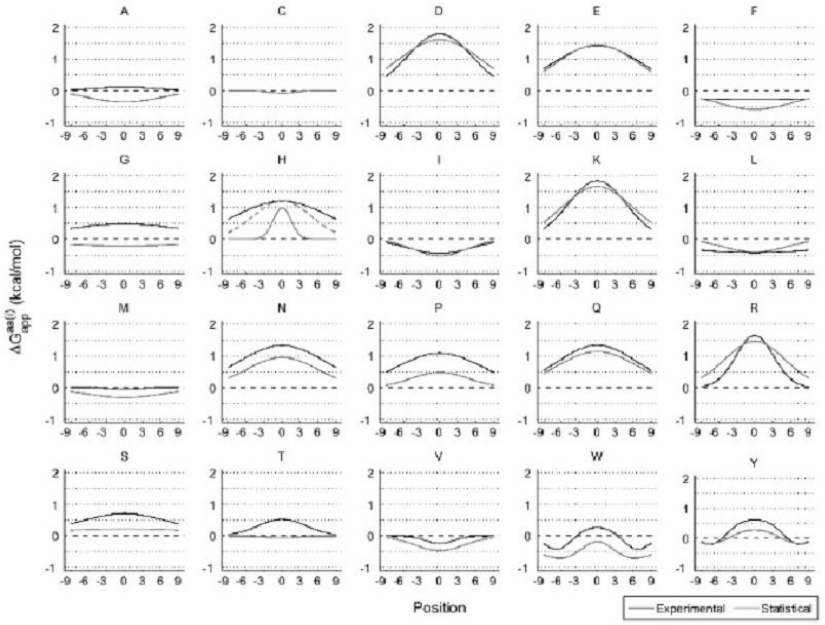


Figure 2: Profiles describing the position-specific amino acid contributions to TM helix insertion efficiency.

3 Prediction of Transmembrane Topology from First Principles

The position-specific scale of amino acid contributions shown in Fig. 2 was further implemented in a simple hidden Markov model-based method to predict TM topology. In a benchmark on known 3D structures of membrane proteins, our simple method was found to perform on par with the current best statistics based TM topology predictors, which often contain hundreds of parameters optimized on known membrane protein topologies, Table 1. TM helices containing e.g. charged residues towards the interfacial region that are missed by other methods, can typically be found using the position-specific information inherent in our method.

First principles-based methods	Single	Multi
SCAMPI	76%	85%
TopPred$^{\Delta G}$	79%	83%
TopPred II	70%	-

Statistics-based methods	Single	Multi
MEMSAT/MEMSAT3	57%	85%
TMHMM/PRODIV-TMHMM	60%	80%
HMMTOP	73%	74%
Phobius/PolyPhobius	63%	66%

Table 1: Fraction of correctly predicted topologies for single- and multiple-sequence versions of different prediction methods on a high-resolution benchmark set of 123 PDB chains. SCAMPI and TopPred$^{\Delta G}$ are based on the profiles in Fig. 2.

References

[1] Hessa, T., Meindl-Beinker, N.M., Bernsel, A., Kim, H., Sato, Y., Lerch-Bader, M., Nilsson, I., White, S.H. and von Heijne, G. Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature*, 450:1026–1030, 2007.

BACOLAP: BAC Assembly based on Bit-Vectors

Jens-Uwe Krause,[1] Juergen Kleffe[2]

1 Introduction

Genome projects calculate and publish overlapping pairs of BACs in order to derive longer genomic sequence contigs sometimes called super BACs. These computations are expensive. A single dynamic programming semi-global alignment of two BACs can take more than 10 hours. Hence, the large number of such alignments required for genome projects make this method impractical. Faster algorithms for assembling reads and ESTs, such as CAP3 [3] and the TGI clustering tool [6], are all unable to handle BAC size sequences ranging from 100 to 300 kb. Other fast programs based on heuristic algorithms for local sequence comparison such as the BLAST version BL2SEQ [7] and ClustDB [4] can deal with large sequence sizes but often fail to detect complete overlaps since increased local error rates cause early termination of local alignments. Then additional analysis is required to select and combine numbers of local matches to complete sequence overlaps. Based on the SeqAn C++ library [9] we therefore developed a faster and more direct tool to solve this problem.

2 Methods

Assuming two BACs A and B to overlap as shown in Figure 1 our new program BACOLAP uses a combination of Myers [5] linear time bit-vector algorithm and Ukkonens [8] cut-off heuristic for approximate string matching to compare the first 300 characters of sequence B with all of sequence A using edit distance in order to detect potential start positions of the overlap in sequence A. Next, using the same method, the last 300 characters of sequence A are compared with a limited region of sequence B in order to identify possible ends of sequence overlaps. Note that an assumed maximal error rate limits the length difference of the two overlapping sections of sequences A and B. Finally a combination of Myers [5] bit-vector algorithm with the divide and conquer method by Hirschberg [2] quickly calculates the exact edit distance alignment for given start and end positions in sequences A and B, respectively. The latter algorithm was developed by Aiche et al. [1]. Not seldom BACOLAP generates more than one possible sequence overlap with error rate below some given threshold. Then a special post processing of the alternative alignments reveals repeats, low complexity regions, possible sequencing errors and clone differences within the overlapping sequence sections.

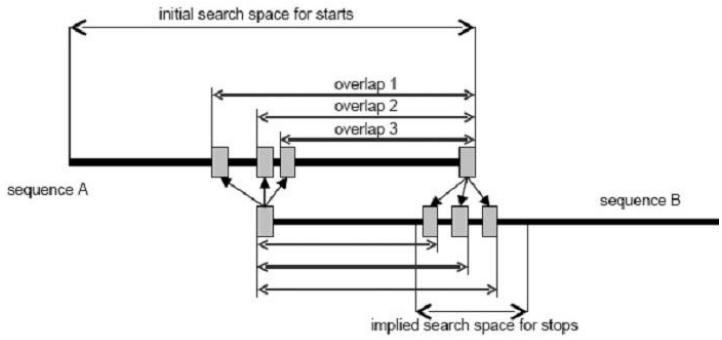


Figure 1: Potential overlaps for sequence A and B.

[1]Department of Bioinformatics, University of Leipzig, Germany. Email: jensenuk@web.de
[2]Institute for Molecularbiology and Bioinformatics, Charit'e, Berlin, Germany. Email: juergen.kleffe@charite.de

3 Results

The performance of BACOLAP was tested by calculating the overlaps of 954 completely sequenced BAC pairs published by the medicago sequencing consortium [10]. The maximum alignment error rate was set to 33.3%. For comparison we ran ClustDB [4] with a maximum error rate of 33.3% within each alignment window of size 300 and BL2SEQ [7] with gap cost –2, gap extension cost –1, mismatch cost –1 and match cost 1. As shown in Table 1, BACOLAP confirmed 944 overlaps and suggests multiple solutions in 75 of these cases. ClustDB [4] confirmed 906 and BL2SEQ confirmed only 874 proving maximal sensitivity for BACOLAP. The increased time of computation is acceptable. 38 overlaps found by BACOLAP but not found by ClustDB contain insertions into one sequence which cause inhomogeneous alignment quality. The window alignment criterion of ClustDB was developed to detect such cases. The local alignment tool BL2SEQ got irretated for the same reason. In other 32 cases BL2SEQ found local matches reaching close to sequence ends. The 10 overlaps missed by BACOLAP were missed by BL2SEQ and ClustDB as well. For 5 of these cases we observed, that no overlap could be found because there is a sequence contamination at the beginning or the end of one of the sequences. For the other 5 cases BL2SEQ could not find any local alignment that could belong to an overlapping region. We also observed that the overlaps found by all considered programs were mostly the same. This suggests to use BACOLAP only on those pairs of BACs for which ClustDB or BL2SEQ failed. Such a combined procedure takes only 26 minutes to confirm 944 overlapping BACs. But an important advantage of BACOLAP is that it finds multiple solutions if possible, while ClustDB and BL2SEQ provide just one. Multiple solutions may suggest further sequencing to correctly determine the considered genomic sequences and give a reason to use BACOLAP for all BAC pairs.

	BACOLAP	ClustDB	BL2SEQ	ClustDB & BACOLAP
found overlapping pairs	944	906	874	
overlaps found with multiple solutions	75	0	0	
overlaps not found	10	48	80	
time elapsed	39 min	19 min	8 min	26 min

Table 1: Results of a comparison between BACOLAP, ClustDB and BLAST 2 Sequences.

References

[1] Aiche, S., Doring, A., Kleffe, J. 2007. Fast and exact global sequence alignment. Poster, *German Conference on Bioinformatics* (GCB2007), Sept. 26–28, Potsdam, Germany.

[2] Hirschberg, D.S. 1975. A linear space Algorithm for computing maximal common subsequences. *Commun. ACM*, 18(6):341–343.

[3] Huang, X., Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res.*, 9:868–877.

[4] Kleffe, J., Moeller, F., Wittig, B. 2007. Simultaneous identification of long similar substrings in large sets of sequences. *BMC Bioinformatics*, 8(Suppl.5):S7.

[5] Myers, G. 1999. A fast bit-vector algorithm for approximative string matching based on dynamic programming. *Journal of the ACM*, 46(3):395–415.

[6] Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., Quackenbush, J. 2002. TIGR Gene Indices Clustering Tools(TGICL): A software system for fast clustering of large EST datasets. *Bioinformatics*, 19:651–652.

[7] Tatusova, A.T., Madden, T.L. 1999. Blast 2 sequences—a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett.*, 174:247–250.

[8] Ukkonen, E. 1985. Algorithms for approximate string matching. *Information and Control*, 64:100–118.

[9] `www.seqan.de`.

[10] `www.medicago.org`.

Host Pathogenesis and Lateral Gene Transfer Revisited: Challenges to Evolution

Rohit Reja,[1] VenkataKrishnan AJ,[1] Sandeep Kumar Yelakanti,[2] Vishal Kumar Nangala,[2] Umesh Roy,[2] Rajib Bandopadhyay,[3] Prashanth Suravajhala[4]

1 Abstract

Bacteria are simple but make complexity possible through various diseases that they cause. Various pathogenic bacteria have proven to be important not only in terms of pathogenesis but also in the light of evolution of any genes that might have transcended from bacteria to the host through lateral gene transfer. Here we bring a short over view challenging the role of gene transfer in pathogenic and non pathogenic bacteria that take light in evolution.

2 Background and Motivation

So far, there are over 874 micro-organisms known to be sequenced while many of the bacterial proteins that are involved in host pathogenesis might have transferred to primate hosts like human. It has earlier been stated that the estimates of bacterial diversity from various sources show that pathogens represent a very small portion of microbial species while most of them don't cause infections. With Systems Biology on rise, researchers are trying to understand Protein-protein interaction (PPI) networks, if such transfer is eminent between organisms using interologs, allowing model organisms like bacteria to supplement the interactomes of higher eukaryotes. It may be noted that the pathogenic bacteria might have lost some virulent genes while non-pathogenic ones might have acquired virulence factors during evolution. Of late, there are experiments showing such hypothesis that have taken two general forms: in the first, genes from the pathogen are assayed for their ability to confer a virulence phenotype upon a normally avirulent strain, and in the second, genes from pathogens are also tested via mutational analysis for their role in virulence. Analysis of sequences recovered by these methods has made it evident that many of the genes required for virulence are restricted to pathogenic organisms and have been introduced into genomes by lateral transfer. Although point mutations may sometimes modulate a virulence phenotype, gene acquisition is much more prevalent as the basis for virulence evolution within lineages. This process is so persistent that species-specific chromosomal regions containing virulence genes are now classed under the general heading of "pathogenicity islands". Our bioinformatical approach aims at identifying proteins or genes involved in host pathogenesis considering *Streptococcus pyogenes* strains as an example.

3 Conclusions

We conclude and foresee the following challenges for identifying proteins or genes involved in host pathogenesis:

1. Finding substitution rates within a pair of protein homologues in different strains. In doing so, several housekeeping loci could be chosen for the characterization of isolates thereby determining their population genetic structure. The nucleotide sequence can then be determined for changes of substitutions.

2. As the evolutionary transitions underlying pathogenic and non-pathogenic means are varied, it has been understood that most necessitate gene transfer and gene loss. We could come across

[1]Vellore Institute of Technology, India.
[2]Precision Biotech, India.
[3]Department of Biotechnology, Birla Institute of Technology, Mesra, Ranchi-835215, Jharkhand, India.
[4]Department of Science, Systems and Models, Roskilde University, Denmark.

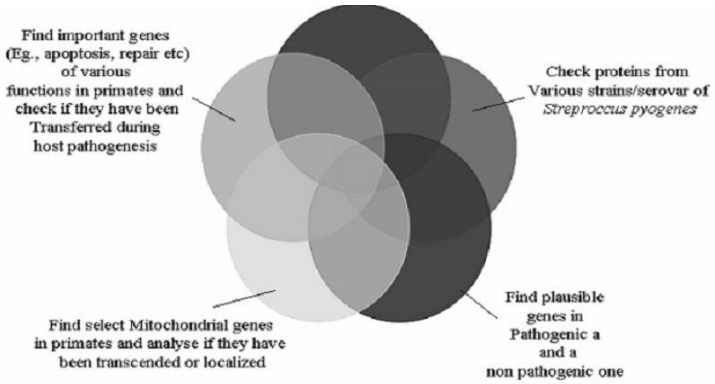


Figure 1: Acquiring genes necessary for host interactions is known to be the most successful methods to identify if virulence is associated with any of the non pathogenic strains or vice versa. The left main Venn indicate the strains with point mutations while the right main Venn denote the virulence phenotype of the strain in discussion (Streptococcus spp). The "Pathogenecity Islands" take up the union of these two strain sets resulting in various permutations and combinations of strains. The other overlapping Venn determine the proteins that possibly could play the role in host-pathogenesis. Finding genes involved in various functions could understandably determine how many genes evolute or devolute, genes lost and found.

bacterial lineages from *Rickettsia prowazekii* as it is considered to be the mitochondrial progenitor. As more and fuller sequences are available after whole genome shotgun (WGS) approach, there is a tremendous scope in developing tools that bacterial proteins have roots traced to mitochondrial proteins of hosts.

3. A need for a database of such proteins could essentially be developed that would be of wide use for evolutionary biologists involved in understanding molecular phylogeny.

References

[1] Aravind L., R. L. Tatusov, Y. I. Wolf, D. R. Walker, E. V. Koonin. *Trends Genet.*, 14:442, 1998.

[2] Brown KR, Jurisica I. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.*, 8(5):R95, 2007.

[3] Chiang-Ni C, Wang CH, Tsai PJ, Chuang WJ, Lin YS, Lin MT, Liu CC, Wu JJ. Streptococcal pyrogenic exotoxin B causes mitochondria damage to polymorphonuclear cells preventing phagocytosis of group A streptococcus. *Med Microbiol Immunol.*, 195(2):55–63, 2006.

[4] Cywes Bentley C, Hakansson A, Christianson J, Wessels MR. Extracellular group A Streptococcus induces keratinocyte apoptosis by dysregulating calcium signalling. *Cell Microbiol.*, 7(7):945–955, 2005.

[5] Delong E. F. *Trends Biotechnol.*, 15:203, 1997.

[6] Gray M. W., G. Burger, B. F. Lang. *Science*, 283:1476, 1999.

[7] Mark C. Enright, Brian G. Spratt, Awdhesh Kalia, John H. Cross, and Debra E. Bessen. Multilocus sequence typing of Streptococcus pyogenes and the relationships between emm Type and Clone. *Infect Immun.*, 69(4):2416–2427, 2001.

[8] Nelson K. E., et al., *Nature*, 399:323, 1999.

[9] Web References: (1) Evolutionary dynamics of disease-related genes, 8 October 2002, `http://www.mscs.mu.edu/~cstruble`. (2) `http://en.wikipedia.org/wiki/Necrotizing_fasciitis`.

ATGC-Dom: Alignment, Tree, and Graph for Comparative Proteomes by Domain Architecture

Duangdao Wichadakul,[1] Supawadee Ingsriswang, Eakasit Pacharawongsakda, Boonyarat Phadermrod, Sunai Yokwai

1 Introduction

Protein domains are units of evolution [11, 13]. Domain combination analysis has been applied for examining proteins in various aspects. For instances, the analysis of co-occurring domains were related to protein functions [1, 6] and the prediction of protein cellular localization [10, 3]. Domain fusion was used for predicting protein-protein interactions [9, 4]. Domain graph [15] and domain distance [2] were introduced for exploring global properties of proteins in the genomes and investigating protein evolution, respectively. While various analyses of protein domains have been performed, available Web-based tools and servers such as PDART [8], CDART [5], and PfamAlyzer [7] mainly enable protein homology search by domain architectures (DAs). ATGC-Dom Web server was built with the aim to enable the comprehensive and customizable comparative analysis of proteomes based on DAs. It integrates three main analyses: (1) comparative proteomes based on DA search and alignment, (2) comparative domain versatilities and abundances based on domain graph, and (3) comparative protein evolutions based on domain distance. For customizable analyses, the user could either provide their own data sets in InterProScan raw format for various domain prediction tools or select data sets from system-provided database. We describe the three main features of the ATGC-Dom Web server in the following.

2 ATGC-Dom Web Server Features

2.1 Comparative Proteome

The "comparative proteome" page lets the user enter proteins of interest and search for target proteins with the same or similar domain architectures. Proteins of interest could be provided by the user in raw format of InterProScan result. Or, the user could search for proteins of interest from system-provided database using (1) general terms such as "flowering", "circadian rhythm", or Gene Ontology (GO) ID such as "GO:0007623", or (2) a combination of arbitrary domains. Proteins of interest will be searched against target proteins by which the user provides as the other InterProScan result in raw format or the user selects from system-provided database. The user may also specify DA score for the cutoff. The search result is in a BLAST-like fashion summarizing number of matched target proteins by target organisms for each protein of interest. The user may explore the alignments of the matches in details. The results of comparative proteome highlight the conservation and diversification of proteins of interest based on their domain architectures within and across input data sets. They suggest protein sets with possibly redundant functions, possible annotations for unknown proteins, single copy genes in the genome, etc.

2.2 Comparative Domain Versatility and Abundance

The "comparative domain versatility and abundance"[2] page lets the user explore versatility and abundance of protein domains within and among protein sets (e.g. among pathways in the same organisms, or among organisms for the same pathway). The user may provide some protein sets as InterProScan resulted files in raw format and select other sets from the system-provided database. The search result is in a table fashion summarizing the versatility and abundance of each protein domains for each protein set. The table allows the user to sort protein domains according to their versatilities or abundances. The user may explore the domain graph and protein lists of each co-occurring domains in a protein set and

[1]Information Systems Laboratory, National Center of Genetic Engineering and Biotechnology (BIOTEC), Pathumthani, Thailand 12120. Email: duangdao.wic@biotec.or.th

[2]A domain versatility represents the number of distinct domains that could be co-occurring with a considered domain. A domain co-present abundance represents the number of proteins in which the domain and its co-occurring domains are present [14].

compare domain graphs among protein sets. Also, the domain graph is customizable to have direction, where an arrow from domain A to domain B represents the having of proteins with two consecutive domains A and B in the order from N- to C- terminals. The user may export domain graph in JPG, PNG, SVG, or PDF format. The domain graphs visualize conserved and diverged co-occurring domains across input data sets with different versatilities and abundances.

2.3 Comparative Protein Evolution

The "comparative protein evolution" page appears to the user after the user chooses all or some of the proteins resulted from other analyses. It calculates a distance matrix according to distances of domain architecture alignments. The user may choose to compare trees built from (1) different algorithms, or (2) different search tools (e.g. hmmpfam, hmmsmart, etc.), as well as (3) DA-based and sequence-based distance matrixes. The user may interactively explore trees of the proteins of interest and their domain architectures in scalable vector graphics (SVG) images. Proteins from different organisms are differentiated by colors. The user may export the images in JPG, PNG, SVG, or PDF formats. In addition, we incorporated a software tool for phylogeny comparison [12] for the user to interactively compare trees. The comparative protein evolution results help a user to explore common ancestors, conserved domains among proteins during the evolution, and lineage-specific domain architectures

3 Concluding Remarks

The ATGC-Dom Web server provides a comprehensive comparative analyses of proteomes based on domain architectures. It integrates aspects of domain architecture analyses into an all-in-one software suite publicly accessible by users via web interfaces. It is designed as a generic system, where user-provided data sets are allowed. Moreover, it is distinguishable from previous systems where it not only performs protein search and alignment but also allows the exploration of protein evolution, as well as helps examining versatile domains and their compositions.

References

[1] Bashton, M. and Chothia, C. 2002. The geometry of domain combination in proteins. *Journal of Molecular Biology*, 315:927.

[2] Bjorklund,A.K., Ekman, D., Light, S., Frey-Skott, J. and Elofsson, A. 2005. Domain Rearrangements in Protein Evolution. *Journal of Molecular Biology*, 353:911.

[3] Bork, P., Hofmann, K., Bucher, P., Neuwald, A. F., Altschul, S. F. and Koonin, E. V. 1997. A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. *FASEB J.*, 11:68–76.

[4] Enright, A. J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C. A. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:6757.

[5] Geer, L. Y., Domrachev, M., Lipman, D. J. and Bryant, S.H. 2002. CDART: Protein Homology by Domain Architecture. *Genome Res.*, 12:1619–1623.

[6] Hegyi, H. and Gerstein, M. 2001. Annotation Transfer for Genomics: Measuring Functional Divergence in Multi-Domain Proteins. *Genome Res.*, 11:1632–1640.

[7] Hollich V. and Sonnhammer, E. L. L. 2007. PfamAlyzer: domain-centric homology search. *Bioinformatics*, 23:3382–3383.

[8] Lin, K., Zhu, L. and Zhang, D.-Y. 2006. An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics*, 22:2081–2086.

[9] Marcotte, E. M., Pellegrini, M., Ng, H.-L., Rice, D.W., Yeates, T. O. and Eisenberg, D. 1999. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science*, 285:751–753.

[10] Mott, R., Schultz, J., Bork, P. and Ponting, C. P. 2002. Predicting Protein Cellular Localization Using a Domain Projection Method. *Genome Res.*, 12:1168–1174.

[11] Murzin, A. G., Brenner, S. E., Hubbard T. and Chothia, C. 1995. SCOP: A structural classi
cation of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536.

[12] Nye, T. M. W., Lio, P. and Gilks, W. R. 2006. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics*, 22:117–119.

[13] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. 1997. CATH: A hierarchic classification of protein domain structures. *Structure*, 5:1093–1108.

[14] Vogel. C., Teichmann, S.A. and Pereira-Leal, J. 2005. The Relationship Between Domain Duplication and Recombination. *Journal of Molecular Biology*, 346:355.

[15] Ye, Y. and Godzik, A. 2004. Comparative Analysis of Protein Domain Organization. *Genome Res.*, 14:343–353.

Predicting Protein Interactions Using Interacting Motif Pairs

Guimei Liu,[1] Jinyan Li,[2] Suryani Lukman,[3] Limsoon Wong[4]

1 Introduction

High-throughput protein interaction data are becoming the foundation of many biological discoveries. However, high-throughput protein interaction data are often associated with high false positive and false negative rates. We develop here a computational method to identify missing interactions from high-throughput data. Our method generates interacting motif pairs from protein sequences and protein interaction networks, and uses them to predict new protein interactions. A confidence score is calculated for every interacting motif pair, and the interacting motif pairs are then used to assign a confidence score to every protein pair that does not interact in the given interaction network to indicate their possibility of being false negatives. We generated interacting motif pairs using the DIP yeast interaction dataset, and evaluated the predicted protein pairs using functional homogeneity. We showed that the interactions predicted by our method have high functional homogeneity, and 62 of the top 100 predictions can be found in the MIPS CYGD database.

2 Method

Given a protein interaction network, our method works in four steps.

Step 1: Purify interaction network. It has been estimated that more than half of current high-throughput data are spurious. Therefore, we first remove spurious interactions from the protein interaction network using a simple measure called CD-distance, which was shown to be very effective in finding false positive errors from high-throughput interaction data [1]. The CD-distance between two proteins p_1 and p_2 is defined as $CD(p_1, p_2) = 2 * A_{12}/(A_1 + A_2)$, where A_{12} is the number of common interacting partners of p_1 and p_2, and A_1 and A_2 are the number of proteins interacting with p_1 and p_2.

Step 2: Generate motifs. Our motif generation method is similar to that used in [2]. We identify groups of proteins that have common interacting partners, called CP protein groups, from the purified interaction network. A CP protein group contains at least l proteins and have at least k common interacting partners. Then for each group, we find sequence motifs from the associated protein sequences using PROTOMAT. To avoid generating too many highly similar motifs, we consider only the maximal CP protein groups for motif generation.

Step 3: Generate interacting motif pairs. We consider every possible pair of motifs and calculate their interacting confidence scores. The confidence of a motif pair (m_1, m_2) is defined as $conf(m_1, m_2) = N_{interact}(m_1, m_2)/N_{total}(m_1, m_2)$, where $N_{interact}(m_1, m_2)$ is the number of interacting protein pairs containing (m_1, m_2) and $N_{total}(m_1, m_2)$ is the total number of distinct protein pairs containing (m_1, m_2).

Step 4: Assign confidence scores to protein pairs. The confidence of a pair of proteins (p_1, p_2) interacting with each other is defined as $conf(p_1, p_2) = CD(p_1, p_2) * Conf_{mtf}(p_1, p_2)$, where $conf_{mtf}(p_1, p_2)$ is the maximal confidence of the motif pairs contained in (p_1, p_2). After the confidence scores of the protein pairs are calculated, we rank the protein pairs in descending order of their confidence, and predict the top ranked protein pairs to be interacting.

3 Results

In our experiments, we use the DIP yeast interaction dataset dated 12/03/2006. True interacting proteins usually share some common functional role. Hence we use the degree of functional homogeneity among interacting protein pairs as one of the measurements to evaluate our method. We use the molecular functional annotations in Gene Ontology (GO) to calculate functional homogeneity.

[1]School of Computing, National University of Singapore, Email: liugm@comp.nus.edu.sg
[2]School of Computer Engineering, Nanyang Technological University, Email: jyli@ntu.edu.sg
[3]Institute for Infocomm Research, Email: lukman1a@gmail.com
[4]School of Computing, National University of Singapore, Email: wongls@comp.nus.edu.sg

Our method ("motif pairs") uses both interacting motif pairs and the CD-distance of the protein pairs to calculate the confidence scores. We compare it with the method that uses CD-distance alone to assign confidence scores to protein pairs. Figure 1 shows the degree of functional homogeneity of the predicted interactions using the two methods. It shows that the generated interacting motif pairs are effective in finding false negative errors. The degree of functional homogeneity of the top 100 interactions predicted by our method is 74.4%, which is significantly higher than the overall functional homogeneity of the protein pairs not in the DIP yeast interaction dataset (11.5%).

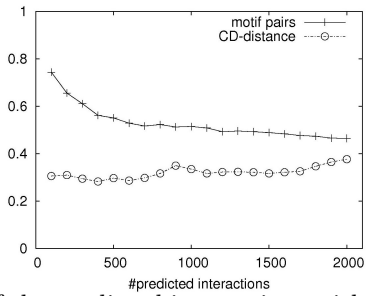
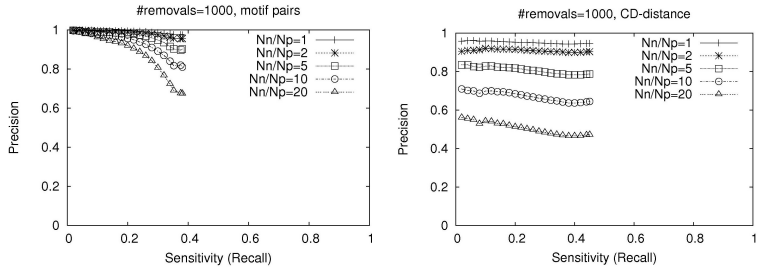


Figure 1: Functional homogeneity of the predicted interactions with respect to the number of predictions

We also verify the predicted interactions in the MIPS CYGD database. For the top 100 interactions predicted by our method, 62 of them are in the MIPS CYGD database. Among the 62 interactions, 5 interactions are direct physical interactions supported by coimmunoprecipitation experiments, one is a genetic interaction, and the remaining 56 interactions are found in complexes. For the top 100 interactions predicted by CD-distance, only 17 are in the MIPS CYGD database, and only one of them is a direct physical interaction. The remaining 16 interactions are found in complexes.

We further study the ability of our method in discerning interacting and non-interacting protein pairs as follows. We randomly remove some interactions from the original DIP yeast interaction dataset, and use the remaining interactions to generate motif pairs. The removed interactions are regarded as true interactions, and they are used to form testing datasets together with some artificially generated non-interacting protein pairs. We generate 100 testing datasets, and we use the average for the final results. Let N_p be the number of interactions removed and N_n be the number of non-interacting protein pairs selected.



(a) motif pairs (b) CD-distance

Figure 2: Effects of the ratio of N_n to N_p (N_p=1000)

Figure 2 shows the performance of the two methods under different ratios of N_n to N_p when 1000 interactions are removed. With the increase of the number of non-interacting protein pairs in the testing datasets, more non-interacting protein pairs are predicted to be interacting, but the number of interacting protein pairs that are predicted to be interacting remains the same, so the precision of both algorithms decreases, and the CD-distance method suffers more than the motif pair method.

References

[1] J. Chen, et al. 2006. Increasing confidence of protein-protein interactomes. In *Proc. 17th International Conference on Genome Informatics*, pp. 284-297.

[2] H. Li, et al. 2006. Discovery motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics*, 22(8):989–996.

Dynamic Programming and Approximation Algorithms for the Simplified Partial Digest Problem

Jacek Blazewicz,[1] Edmund K. Burke,[2] Marta Kasprzak,[3]
Alexandr Kovalev,[4] Mikhail Y. Kovalyov[5]

1 Introduction

Restriction mapping is a common technique in developing a physical genetic map. One of the main approaches in the restriction mapping is the partial digest method. Unfortunately, partial digest suffers from several serious drawbacks, which prevented it from being broadly used in biological laboratories [4]. In order to overcome the disadvantages of the partial digest method, Blazewicz et al. [1] proposed the simplified partial digest method, which is much easier to implement, and more accurate and robust with regard to the experimental errors. It consists of two biochemical experiments. In the first, one enzyme cuts clones of the target DNA at exactly one appropriate restriction site, and in the second, it cuts at all n appropriate sites. The lengths of the obtained DNA fragments are further measured. The modeling combinatorial problem (Simplified Partial Digest Problem, SPDP) is NP-hard in the strong sense [2]. Several solution algorithms for SPDP are known in the literature. They are able to solve moderate size instances of the problem, and do not always provide satisfactory quality in the case of measurement errors. We present new efficient approaches for SPDP. The input for SPDP consists of two multisets of numbers, A and B, where A contains n pairs of end distances, and B contains $n+1$ intersite distances. The output is a sequence of n restriction sites (DNA linear structure) represented as points in the interval $[0, L]$, where L is the length of the target DNA

2 An Overview of New Algorithms for SPDP

For the error-free case of SPDP we present an original dynamic programming algorithm, denoted as DP, which is efficient when the number of distinct intersite distances, denoted as q, is small. The worst case running time of DP is $O(n^{2q})$. Approximation algorithms are derived for optimization versions of SPDP. These algorithms do not guarantee the construction of a correct DNA map but they are able to construct a DNA map, which is sufficiently close to the correct one. Their advantage is the computational time, which is polynomial in the worst case. On the basis of our new graph-theoretic model for SPDP, we develop three practically useful heuristics, denoted as SWITCH, PATH-F and PATH(x), where x is the algorithms parameter. The model itself can be applied to reduce the search space for an optimal (exact) solution of SPDP. The worst case running time of SWITCH is $O(n \log n)$, and those of PATH-F and PATH(x) are $O(n^5)$ and $O(n^2 max\{n, x\})$, respectively.

3 Computational results

We used the simplified partial digest method to find a solution for the incomplete partial digestion data obtained in [3]. An algorithm for the Partial Digest Problem is useless in this case because it does not work with incomplete experimental data. Computational experiments with the proposed exact and approximation algorithms demonstrated their efficiency. Specifically, algorithm DP was able to solve instances of the error-free SPDP with hundreds of restriction sites in less than one second on a standard PC. It outperforms the fastest enumerative algorithm (ENUM) on random data if the number of distinct intersite distances q is less than 10% of the total number of restriction sites n, see Table 1.

[1]Institute of Computing Science, Poznan University of Technology, and Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland. Email: jblazewicz@cs.put.poznan.pl

[2]Automated Scheduling Optimization and Planning Group, School of CSiT, University of Nottingham, Nottingham NG8 1BB, UK. Email: ekb@cs.nott.ac.uk

[3]Institute of Computing Science, Poznan University of Technology, and Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland. Email: marta@cs.put.poznan.pl

[4]Institute of Computing Science, Poznan University of Technology, Poznan, Poland. Email: akovalev@cs.put.poznan.pl

[5]Faculty of Economics, Belarusian State University, and United Institute of Informatics Problems, National Academy of Sciences of Belarus, Minsk, Belarus. Email: koval@newman.bas-net.by

	Running time, msec			
	ENUM		DP	
q	Average	Std deviation	Average	Std deviation
40	106	43.1	75	24.8
50	111	28.1	85.2	18.5
60	125	66	97.8	43.8
70	114	41.8	96.3	32.9
80	140	42.7	126	38.7
90	151	137	140	106
100	162	89.6	161	87.3
110	154	114	161	108
120	175	92.4	196	103
130	152	144	193	168
140	190	226	241	272
150	174	115	282	208

Table 1: Running time of algorithms ENUM and DP. Random data, n = 1000.

Computational experiments with our approximation algorithms demonstrated that, for instances of SPDP generated using real DNA sequences from GenBank with $20 \leq n \leq 50$, and for randomly generated instances with $n = 50$, algorithms PATH-F and PATH(n^2) always delivered an exact solution, see entries "100%" in Figure 1. The fastest of our approximation algorithms is SWITCH. It securely produces all correct end distances and at least $n/2 + 1$ correct intersite distances. For $20 \leq n \leq 50$ and data from GenBank, it found 90% of exact solutions, and the average percentage of correct intersite distances was 99.9%.

Size of	Average %*Exact*		
instances	SWITCH	PATH-F	PATH(n^2)
$20 \leq n \leq 50$	90%	100%	100%
$50 \leq n \leq 100$	58.3%	98.7%	93.3%
$100 \leq n \leq 150$	12.5%	84.5%	51.3%
$150 \leq n \leq 200$	0%	49.6%	0.32%
$200 \leq n \leq 250$	0%	13.8%	0%
$250 \leq n \leq 300$	0%	0%	0%
$300 \leq n \leq 350$	0%	0%	0%
$350 \leq n \leq 400$	0%	-	0%

(a) Genbank data

	Average %*Exact*		
n	SWITCH	PATH-F	PATH(n^2)
50	58%	100%	100%
100	10%	100%	96%
200	0%	56%	0%
300	0%	14%	0%
400	0%	0%	0%
500	0%	0%	0%
600	0%	-	0%

(b) Random data

Figure 1: Percentage of exact solutions found on Genbank data and random data.

All three proposed approximation algorithms can be considered as a good tool for solving real-life SPDP instances. The choice of the specific approximation algorithm depends on the input data, desired solution quality, and computational time limit.

References

[1] Blazewicz, J., Formanowicz, P., Kasprzak, M., Jaroszewski, M., and Markiewicz, W.T. 2001. Construction of DNA restriction maps based on a simplified experiment. *Bioinformatics*, 17:398–404.

[2] Blazewicz, J. and Kasprzak, M. 2005. Combinatorial optimization in DNA mapping: A computational thread of the simplified partial digest problem. *RAIRO—Operations Research*, 39:227–241.

[3] Dudez, A., Chaillou, S., Hissler, L., Stentz, R., Champomier-Verges, M., Alpert, C. and Zagorec, M. 2002. Physical and genetic map of the lactobacillus sakei 23k chromosome. *Microbiology*, 148:421–431.

[4] Pevzner, P.A. 2000. *Computational Molecular Biology: An Algorithmic Approach*. MIT Press, Cambridge, MA.

Information-Theoretic Analysis for Exploring Cell Death-Survival Signaling

Yew Chung Tang,[1] Gregory Stephanopoulos,[2] Heng-Phon Too[3]

1 Introduction

Living cells have a remarkable ability to process information generated by extracellular stimuli and make complex behavioral decisions. Many of these cell fate decisions, including death or survival, are made by networks of dynamically regulated signaling proteins. Deregulation of signaling networks governing these cellular mechanisms results in the invasive nature and uncontrolled growth of tumor cells that are hallmarks of cancer. To effectively target signaling mechanisms in cancer cells, it is necessary to have a systems-level understanding of cell signaling. We applied a systems biology approach to the study of the platelet-derived growth factor (PDGF) signalling network and its protective effect against cell death in T98G glioblastoma cells initiated by tumour necrosis factor-related apoptosis-inducing ligand (TRAIL).

2 Approach

We implemented an iterative framework for exploring the PDGF signalling network: using information-theoretic analysis of multivariate, dynamic and quantitative signalling data generated from a reduced system of 4 signaling proteins involved in PDGF signalling, we hope to generate novel insights and hypotheses for further exploration of the network. We seek to answer two key questions in understanding cell signaling systems: what are the components involved and how are these components related? We use information-theoretic measures to identify the most informative signals or combination of signals that can predict death-survival responses and further infer causal links between these signals that are important in determining cellular response.

3 Results

Our analysis led to predictions that were experimentally verified using pharmacological inhibition of the signaling network and also generated new hypotheses that are the subject of further investigation.

[1]Singapore-MIT Alliance. Email: tangyc@nus.edu.sg
[2]Department of Chemical Engineering, Massachusetts Institute of Technology. Email: gregstep@mit.edu
[3]Department of Biochemistry, National University of Singapore. Email: bchtoohp@nus.edu.sg

Constructing Transfer Pathways in Multidimensional NMR Spectra of RNAs

Marta Szachniuk,[1] Mariusz Popenda,[2] Lukasz Popenda,[3]
Jacek Blazewicz[4]

1 Introduction

RNA is known to be the foreground actor in the storage and communication of biological data. Recent discoveries has proved that it also performs enzymatic catalysis, thus, being likely to have been an initiator of the first living systems [3] The details of RNA structure influence the functionality of this molecule. Thus, determination of RNA tertiary structure is a crucial task in biological studies, with major contribution from NMR and X-ray crystallography. Despite the progress in both of these techniques, recognition and analysis of RNA structure is still very difficult. In case of NMR spectroscopy, a resonance assignment step is a bottleneck in the process of high-resolution structure determination and makes an analysis of large structures hardly possible. During this procedure transfer pathways between interacting atoms are reconstructed in the spectra and NMR signals are assigned to appropriate atoms of the molecule. The following interactions can be considered: correlation between (i) base-pair protons, (ii) H1 and H6/H8 protons, (iii) ribose protons, (iv) ribose and ^{31}P [5]. Figure 1 shows an example of transfer pathway reconstructed during an analysis of H1 and H6/H8 protons in the fragment of 2D and 3D NMR spectrum obtained for r(ACGU) molecule.

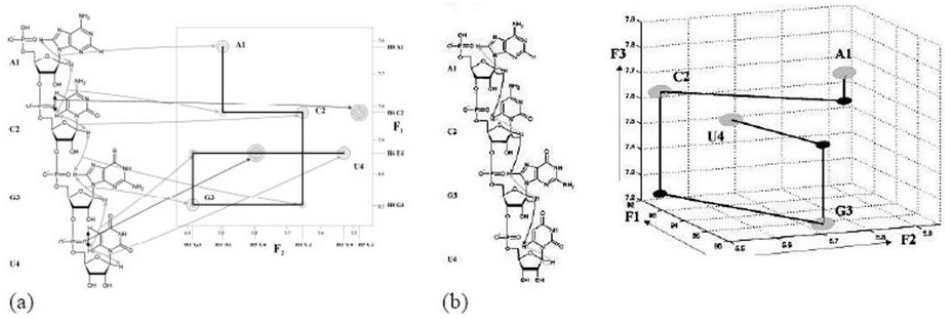


Figure 1: H1-H6/H8 path in 2D NOESY spectrum (a) and C1-H1-H6/H8 path in 3D HSQC-NOESY spectrum (b).

At present, most NMR studies of RNA are based on 2D spectra analysis. Performing 3D experiments is more expensive but provides spectra of much better resolution. Here, we propose a new theoretical graph model to represent the problem of the transfer pathway construction in 3D NMR spectra. Considering one selected pathway of magnetization transfer between H1 and H6 / H8 atoms, we compare its reconstruction within 2D and 3D spectra recorded for RNA molecules. We introduce a new algorithm for an automatic generation of paths in three-dimensional spectra and we conclude on the differences between 2D and 3D aspects of resonance assignment.

2 Methods

The problem of resonance assignment within two-dimensional NMR spectra has been modeled on the basis of graph theory and new type of graph, called NOESY graph, has been introduced [1, 2]. It seemed natural to follow the same idea in case of three dimensions. However, since resonance assignment in three-dimensional spectra differs in several aspects from the 2D case [4], transformation of a NOESY graph to make it represent 3D spectra appeared ineffective. So a new graph model, called a spectral graph, has been proposed.

[1]Institute of Computing Science, Poznan University of Technology, Poland. Email: mszachniuk@cs.put.poznan.pl
[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland. Email: marpop@ibch.poznan.pl
[3]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland. Email: lpopenda@ibch.poznan.pl
[4]Institute of Computing Science, Poznan University of Technology, Poland. Email: jblazewicz@cs.put.poznan.pl

Let us consider undirected graph $G = (V, E)$, where V is a set of vertices and E is a set of edges. Furthermore, let us assume that G has the following properties:

1. every vertex $v_i \in V$ represents a cross-peak from the spectrum;
2. the number $|V|$ of vertices in graph G is equal to the number of cross-peaks in the spectrum;
3. every edge $e_j \in E$ represents a potential connection between two vertices of V having: (a) exactly two common coordinates (we will call it an ordinary edge), (b) exactly one common coordinate (we will call it a diagonal edge);
4. every edge $e_j \in E$ is associated with a label $l_j = \{0, 1, 2, 3, 4, 5\}$ in the following manner:

$$l_j(e_j(v_m, v_n)) = \begin{cases} 0 & \text{if } v_m \text{ and } v_n \text{ differ in Z dimension} \\ 1 & \text{if } v_m \text{ and } v_n \text{ differ in Y dimension} \\ 2 & \text{if } v_m \text{ and } v_n \text{ differ in X dimension} \\ 3 & \text{if } v_m \text{ and } v_n \text{ differ in X, Y dimensions} \\ 4 & \text{if } v_m \text{ and } v_n \text{ differ in X, Z dimensions} \\ 5 & \text{if } v_m \text{ and } v_n \text{ differ in Y, Z dimensions} \end{cases}$$

5. the number $|E|$ of edges in graph G equals all possible ordinary and diagonal connections that can be drawn in the spectrum.

If we consider labels as colors, we obtain 6-edge-coloured graph, i.e. sextuple $G = (V, E_0, E_1, E_2, E_3, E_4, E_5)$, where V is a set of vertices and $E_0 - -E_5$ are disjunctive sets of edges labeled 0–5, respectively. Any 3D NMR spectrum can be transformed into such a graph. Appropriate modification of properties (3) and (4) of the above definition makes it suitable for a representation of n-dimensional spectra, where n =2, 4, 5, etc.

A transfer pathway P_G, which results in a resonance assignment of selected atoms, can be reconstructed within the spectral graph G according to the following rules: every vertex $v \in V$ and every edge $e \in E$ may occur in the path at most once, the path does not contain collinear edges, every edge $e_k \in P_G, k = 1 \ldots l$, with label l_k, satisfies the following condition: $l_k \in \{0, 1, 2\}$, $l_k \neq l_{k+1}$, $lk \neq l_{k+2}$ in case of path crossing homonuclear interactions or $(l_k \bmod 3) = (l_{k+1} \bmod 3)$ in case of heteronuclear interactions. The final condition determines which labels (colors) are considered when the path is constructed.

3 Results

A new algorithm for a reconstruction of transfer pathways within 3D NMR spectra has been designed. It is based on a Hamiltonian path construction procedure and uses domain expert knowledge to decide on the type of a pathway (homo- or heteronuclear interactions considered) and to reduce the search space. The algorithm has been implemented in C language and tested on a set of simulated 3D NMR spectra. As in case of 2D NMR spectra analysis [1, 2], a possibility of an automatic generation of transfer pathways greatly facilitates the process of resonance assignment.

Acknowledgement. This research was supported by a grant of the Polish Ministry of Science and Higher Education.

References

[1] Adamiak, R.W., Blazewicz, J., Formanowicz, P., Gdaniec, Z., Kasprzak, M., Popenda, M. And Szachniuk, M. 2004. An algorithm for an automatic NOE pathways analysis in 2D NMR spectra of RNA duplexes. *Journal of Computational Biology*, 11(1):163–180.

[2] Blazewicz, J., Szachniuk, M. and Wojtowicz, A. 2005. RNA tertiary structure determination: NOE pathway construction by tabu search. *Bioinformatics*, 21(10):2356–2361.

[3] Joyce, G.F. 2002. The Antiquity of RNA-Based Evolution. *Nature*, 418:214–221.

[4] Szachniuk, M., Popenda, M., Klemczak, S. and Blazewicz, J. 2007. An analysis of three-dimensional NMR spectra in RNA structure determination process. *Research report RA-001/2007*, Poznan Supercomputing and Networking Center.

[5] Wuthrich, K. 1986. *NMR of Proteins and Nucleic Acids*. New York: John Wiley & Sons.

Robust Optimization for Biological Network Calibration

Bo Kim,[1,2] Bruce Tidor,[2,3,4] Jacob White[1,2]

As biological systems are being increasingly investigated from the network point of view, there is an escalated demand for computational models that quantitatively characterize those systems. For instance, as dysregulation of apoptosis is found to contribute to various autoimmune diseases and cancer [1], developing comprehensive and predictive models of the signaling pathways for apoptosis may help quantify the effectiveness of candidate treatment targets.

An essential yet time-consuming task in building these models is using available data to calibrate the parameters that define the model. When signaling pathways are modeled using differential equations derived from chemical kinetics, the parameters subject to calibration are chemical reaction rates or initial concentrations of species. The task then is to determine the set of parameter values so that the model generates outputs that match experimental measurements. A major barrier to successful model calibration is the limited amount of available experimental data. Therefore, it is often the case that multiple sets of parameters produce outputs that match the measured data, and it is difficult to determine which of these many parameter sets correspond to a model that will be predictive.

Because many parameters must be estimated from only a small amount of data, additional biologically reasonable constraints that keep the estimation problem tractable may be especially useful in order to select the correct parameter set among many. Based on the intuition that a critical behavior of a system is not likely to have been designed to react dramatically to ubiquitous noise and varying surrounding conditions that cause small parameter changes, robust optimization methods are explored to calibrate computational models of biological systems. Results have been obtained from using an algorithm based on sampling, and they suggest that robustness may be a reasonable biological constraint to add in optimizing for the parameters. Furthermore, including robustness as a constraint seems to make calibration based on noisy measured data more manageable, while indicating the need to carefully examine the choice of data that parameter estimation is performed with respect to.

Research thus far has been conducted primarily in the context of signal transduction pathways, including the mitogen-activated protein kinase, Fas signaling, and epidermal growth factor receptor (EGFR) pathways. In each case, the number of parameters to be estimated was taken to be approximately 2 to 5 times greater than the number of available data points. These pathways have long been under heavy investigation for their relevance to cancer research; in particular, components of the EGFR pathway have been targeted for cancer therapy, with malignant cells of multiple myeloma patients being found to over-express a number of EGFRs and their ligands [2].

References

[1] Hua, F. *et al.* 2005. Effects of Bcl-2 levels on Fas signaling-induced caspase-3 activation: Molecular genetic tests of computational model predictions. *The Journal of Immunology*, 175:985–995.

[2] Johnston, J. B. *et al.* 2006. Targeting the EGFR pathway for cancer therapy. *Current Medicinal Chemistry*, 29:3483–3492.

[1]Research Laboratory of Electronics, Massachusetts Institute of Technology, MA, USA. Email: `kaede11@mit.edu`
[2]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, MA, USA
[3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, MA, USA
[4]Biological Engineering Division, Massachusetts Institute of Technology, MA, USA

Specificity and Affinity of Protein-Protein Interactions: From a Systems Biology to a Molecular Point of View

Pablo Carbonell,[1] Antonio del Sol[1]

1 Introduction

Protein-protein interactions are crucial to most cellular processes. Hence, knowledge about these interactions is essential for understanding biological functions. Most studies of protein networks rely on the analysis of global characteristics of network topologies, ignoring structural and physico-chemical aspects of each interaction. However, the understanding of the organization and functioning of protein interaction networks requires a combination of network analysis with a detailed characterization of the molecular principles that underlie individual protein associations. Here we combine this two approaches by investigating specificity and affinity of protein-protein associations based on a thermodynamic and structural basis.

2 Results

We compiled a dataset of protein-protein interactions with structural information from the yeast interactome. For each interaction we were able to find a representative structure of the complex. Using the structural information, we identified hub proteins, clustered their binding sites, and classified them into singlish-interface hubs (involving only one interface for binding) and multi-interface hubs (involving more than one interface for binding). Furthermore, we counted the number of nonredundant partners interacting through each interface. Specific binding sites were defined as those binding sites interacting with only one partner, whereas promiscuous binding sites interact with more than one partner.

We calculated hydrophobic patches on the protein hub surfaces [1], seeing a clear tendency for the multiinterface hubs to have on average a larger number of patches distributed over their surfaces. This result suggests that the patches analysis could be used to identify multi-interface hubs, which, according to previous studies, are more likely essential for cellular viability [2].

The affinity of the interactions was estimated by calculating the binding free energy of each interaction. Our estimations were based on the rigid structure of the complex, and therefore large entropic contributions such as ordering of disordered binding sites upon binding were not considered. Consequently, in our estimations those interactions associated with disordered binding sites were found to require a high enthalpic contribution to compensate the entropic cost of binding. Based on this finding, and in order to estimate the binding free energy more precisely, we restricted our affinity analysis to protein-protein interactions involving ordered binding sites

We identified residues essential for binding (hotspots) for each protein interface [3], and analyzed how they were distributed across the set of different interactions. Hotspots that were found in more than one interaction might play an essential role for determining affinity, whereas those associated with single interactions could be responsible for binding site specificity.

Specificity and affinity in protein-protein interactions. We observed some significant correlation between the binding free energy per residue (as an estimator of binding affinity) and the number of interacting partners for each binding site (see Figure 1). In other words, promiscuous binding sites tend to interact with lower affinities with their partners in comparison with specific binding sites. Randomization of our data was used to test the significance of our results. Further results showed that, as it was expected from this finding, interactions involving specific binding sites in both interacting partners tend to be stronger than interactions involving promiscuous binding sites in both interacting partners.

[1]Bioinformatics Research Unit, Research and Development Division, Fujirebio, Tokyo, Japan. Email: {pl-carbonell,ao-mesa}@fujirebio.co.jp

3 Conclusion

Protein-protein interactions can be analyzed from a broader perspective by combining network and structural properties. Our results suggest that physico-chemical and thermodynamic properties estimated from structural information such as surface patches, binding hotspots, and free binding energy; might be related to network properties such as binding site specificity. Namely, our analysis on hubs binding sites shows that interaction affinity is regulated by binding specificity. Indeed, promiscuous binding sites are mainly involved in weak interactions, which are common in molecular functions involving several interacting partners, such as transcription regulator activities. Furthermore, we find that interactions associated with disordered binding sites require a high enthalpic contribution to compensate the entropic cost of binding. These results shed light on the mechanism of protein-protein interactions, and may be useful in the discovery process of high-affinity specific compounds that target protein-protein interactions.

References

[1] Jones, S. and Thornton, J.M. 1997. Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.*, 272(1):121–132.

[2] Kim, P. M., Lu, L. J., Xia, Y., and Gerstein, M. B. 2006. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 314(5807):1938–1941.

[3] Ofran, Y. and Rost, B. 2007. Protein-protein interaction hotspots carved into sequences. *PLoS Comput. Biol.*, 3(7):e119.

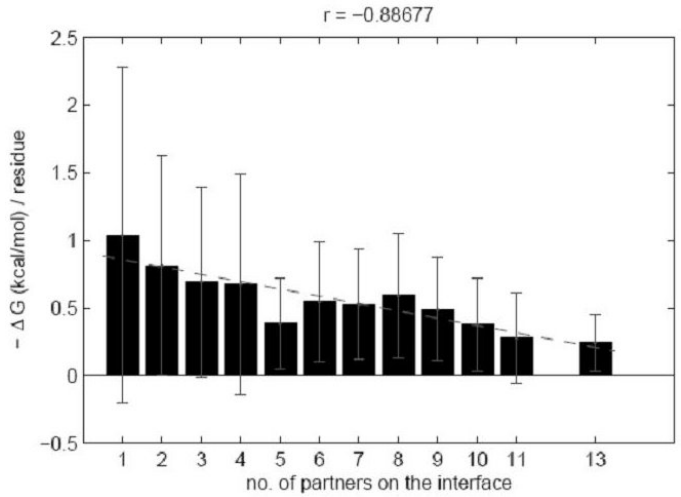


Figure 1: Specificity vs. affinity in protein-protein interactions. Specificity for each protein-protein interface is given by the number of partners which interact through it; affinity is calculated as the average binding free energy per residue for each degree of interface specificity. Error bars have been represented on each bin, as well as the fitted regression line (r = -0.89).

Using Gene Expression Analysis for Drug Pathway Identification: An Example on Nasopharyngeal Carcinoma (NPC)

Dong Difeng,[1] Limsoon Wong[2]

1 Introduction

Nasopharyngeal carcinoma (NPC) is a malignant cancer in the head and neck region, with especially high incidence in South China, Southeastern Asia and North Africa. Recently, a cyclin dependent kinase (CDK) inhibitor, CYC202, is studied for its anti-tumor effect in human NPC cells in vitro and in vivo. Results show that both cell lines and patients in the study responded to the drug treatment dierently. To further investigate the drug response, expression of selected genes for apoptosis, cell proliferation and cell cycle regulation were measured during the process of treatment. Our issue is how to identify the reason for the dierent responses in these NPC individuals using the gene expression data. Biological pathway information has long been incorporated into gene expression analysis for the purpose of treatment response understanding. However, the conclusions are usually too general, and hardly sucient for guiding further research. In our current study, we design a drug pathway identification system, the Drug Pathway Decipherer, which identifies genetic regulations in response to drug treatment that are consistent with respect to a given detailed signaling pathway structure. By applying our system to the NPC dataset, we discover that the status of ERK pathway and apoptosis pathway are differently regulated between responders and non-responders both in vitro and in vivo. Our results indicate that the dysregulation of Ras-ERK pathway and PI3K-Akt-NFkB pathway are probably the mechanisms for CYC202-insensitive NPC cells to resist the drug treatment.

2 Method

The Drug Pathway Decipherer consists of 4 partitions distributed on two biological levels. Figure 1 gives the diagram of its workflow. For each user specified signaling pathway, highly consistent genetic regulations are selected and connected into genetic pathways. The significance of these pathways are then evaluated with their expression correlation against the background. P-value and FDR cutoff are used to control the statistical significance of the identified pathways, and pathway status are derived from the expression of genes on pathways and their associated statistics levels.

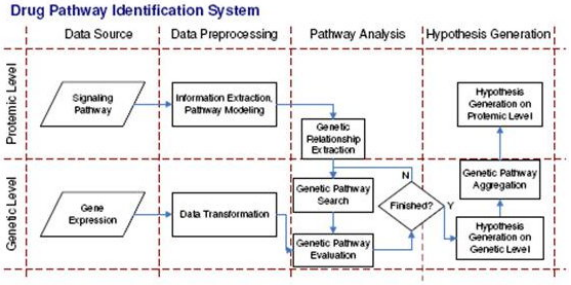


Figure 1: The workflow of the drug pathway identification system.

[1]National University of Singapore, Singapore. Email: dong.difeng@gmail.com
[2]National University of Singapore, Singapore. Email: wongls@comp.nus.edu.sg

3 Results

Figure 2 shows the results of applying our system to NPC cell lines. From the figure, the pathway status of ERK pathway and the apoptosis pathway is obviously differentially regulated in three cell lines, which is also consistent with the observed drug effect; the p-value of this difference reaching reaching 4E-4 for ERK pathway and 2.8E-3 for the apoptosis pathway.

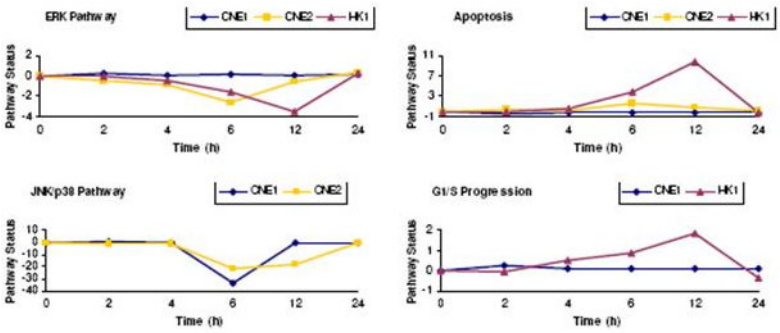


Figure 2: Comparable diagrams of pathway status profiles of the three cell lines.

Table 1 gives the results of pathway status regulation of patients. Patient18 is previously known as a non-tumor sample. The status regulation of ERK pathway and apoptosis pathway of this patient can almost perfectly (with one outlier) separates the other patients into two different treatment response groups, which suggests the regulation of these two pathways is closely related to drug response.

Patient	Response	ERK		JNK/p38		G1/S		Apoptosis	
		Status	Conf.	Status	Conf.	Status	Conf.	Status	Conf.
Pt5	P(ositive)	-2.25	0.98	-3.08	0.99	-	-	1.34	0.99
Pt8	P	-	-	-1.01	0.99	-	-	0.82	0.98
Pt9	P	-0.97	0.98	-	-	0.76	0.95	-	-
Pt14	P	-	-	-	-	-0.61	0.99	-0.86	0.99
Pt16	P	-0.20	0.99	-0.20	0.95	0.29	0.99	1.42	0.97
Pt17	P	-1.02	0.99	-1.02	0.99	-0.33	0.96	1.01	0.99
Pt19	P	-	-	-0.86	0.98	-	-	0.91	0.98
Pt18	No Tumor	**-0.15**	0.99	-	-	0.28	0.99	**0.13**	0.99
Pt1	N(egative)	0.21	0.95	0.52	0.99	1.06	0.97	-1.00	0.98
Pt7	N	-0.10	0.97	-0.68	0.96	0.28	0.98	0.11	0.98
Pt10	N	1.02	0.99	1.16	0.99	-	-	-1.57	0.97
Pt15	N	-	-	-	-	-	-	-1.01	0.98
Pt20	N	1.30	0.98	-	-	-0.93	0.96	-1.68	0.99

Table 1: The results of signaling pathway status estimation for the in vivo dataset: The "response" column shows the molecular response to treatment for patients. The "status" column shows the estimated post-treatment pathway status.

References

[1] Guo,Z. et al. (2007) Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. **Bioinformatics**, 23, 2121-2128.

[2] Ideker,T. et al. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18, s233-s240.

[3] Soh,D. et al. (2007) Enabling more sophisticated gene expression analysis for understanding diseases and optimizing treatments. *ACM SIGKDD Explorations*, 9, 3-14.

[4] Zien,A. et al. (2000) Analysis of gene expression data with pathway scores. *Proceedings of International Conference on Intelligent Systems for Molecular Biology*, 8, 407-417.

Composition of Signaling Pathway Models and its Application to Parameter Estimation

Geoffrey Koh,[1] David Hsu,[2] P.S Thiagarajan[2]

1 Introduction

The functioning of biological pathways depends on the interactions among their constituent elements: genes, proteins, and other molecular species. To gain a systems-level understanding of these complex pathways, we need quantitative models that capture the evolution of such interactions over time. Our focus here is on constructing and, in particular, *composing* dynamic models of signaling pathways.

A biological pathway can be viewed as a network of biochemical reactions. To build a pathway model, we need both the network structure and the parameters – kinetic rate constants, initial conditions, etc. – that govern the individual biochemical reactions. Parameter estimation of a biological pathway model is a challenging problem, due to the high-dimensional search space involved and the lack of accurate data. Furthermore, model construction is an incremental process, due to new players being discovered and additional experimental data on the known players of the pathway becoming available. It is thus important to develop methods for building pathway models that can be easily *refined* and *expanded*.

Conventional parameter estimation algorithms [5] fit pathway parameters to *all* available experimental data. When new data becomes available, the entire procedure is repeated afresh, using both the new and the old data. This wastes significant computation time. More importantly, the old data may not be systematically archived and easily accessible.

We propose to use a probabilistic model known as *factor graphs* [3] to address the above issues. By capturing the local interactions, the factor graph model drastically reduces the search space for parameter estimation. Being a probabilistic model, it also naturally handles noise in the data. Most importantly, it contains multiple parameter estimates encoded as probability distributions rather than a single best estimate. In addition, new experimental data and pathway players can be integrated into the factor graph incrementally.

Both model refinement and expansion rely on a probabilistic inference technique called *belief propagation* [6]. Using this technique, one can propagate local constraints through the entire network and obtain a globally consistent model. Factor graphs have been used to model biological systems [1], but in this earlier work, the main goal is to study the functional correlations among the elements in the pathway rather than the dynamics.

2 Factor Graph Models of Pathway Dynamics

A signaling pathway is a network of biochemical reactions where the reactions are often mediated by enzymes. The dynamics of the pathway is described by a system of ordinary differential equations (ODEs). The ith equation has the form $\dot{\mathbf{x}}_i = f_i(\mathbf{x}(t), \mathbf{p})$, where $\mathbf{x}(t)$ is a vector-valued function describing the concentration levels of molecular species at time t and $\mathbf{p}$ is the set of pathway parameters.

We build a factor graph model for a given system of ODEs. A factor graph is an undirected bipartite graph consisting of *variable nodes* and *factor nodes*. Each variable node corresponds to an unknown parameter or enzyme, and each factor node corresponds to the ODE. The edges of a factor graph represent the dependencies of the reaction rates on the parameter values and the enzyme concentration levels.

We represent each parameter as a probability distribution and associate it with a variable node of the factor graph. For completely unknown parameters, their initial distributions are assumed to be uniform. Other parameters have *a priori* distributions that reflect prior knowledge. These distributions are updated as new data becomes available. Each factor node is associated with a joint probability distribution that captures the dependencies of the factor node on the variable nodes, as specified in the ODEs. We build this distribution by *sampling* the values of the parameters corresponding to the variable nodes involved.

[1]NUS Graduate School for Integrative Sciences and Engineering. Email: g0306427@nus.edu.sg
[2]Department of Computer Science, National University of Singapore. Email: {dyhsu, thiagu}@comp.nus.edu.sg

For each set of sampled parameter values, we simulate the system of ODEs and get a score that is the weighted mean squared difference between the simulated and experimental time-series data. The scores are then normalized to obtain a probability distribution.

3 Pathway Composition and Data Integration

Pathway components can arise in several ways. For instance, in our earlier work [4], we tackled the parameter estimation problem for large pathway models by decomposing them into smaller components. Multiple components can also arise when different pathways – elucidated independently – are linked together. In either case, each pathway component can be represented by its own factor graph. Composing the components then involves "fusing" the corresponding factor graphs at their common variable nodes to form a composite factor graph.

Similarly, we can integrate new data into an existing pathway model represented as a factor graph. We first sample the part of the pathway relevant to the new data and build a new factor graph for this part. We then combine the new and the existing factor graphs by fusing their common variable nodes. This idea is illustrated in Figure 1.

One key issue in composition is to ensure that the local dependencies and constraints in the components are all captured in the composite factor graph and that they are consistent. To achieve this, we use belief propagation to propagate local constraints globalli [6]. Upon convergence, the variable nodes of the factor graph contains the maximum a posteriori distributions of the parameters.

4 Results and Discussion

We tested this approach on a simplified model of the Akt-MAPK signaling pathway [4]. Using four sets of experimental data synthesized on the Akt-MAPK model through simulation, we performed parameter estimation on the model incrementally by adding one data set at a time and applying our composition method. For comparison, we applied two other methods implemented in the modeling software COPASI [2]. All the methods were allocated equal amounts of time for the four data sets. Preliminary results (Figure 2) indicate that our method achieved substantially better estimates.

We are currently extending this work in two directions. Recent experimental developments suggest that cross-talks are common between signaling pathways. By systematically composing pathway models, we plan to construct large signaling pathway models that take into account cross-talks between the individual pathway components. Second, we plan to improve the sampling process for building the joint distributions associated with the factor nodes. Currently we sample uniformly over the entire local parameter space. A "guided sampling" approach can improve the results by focusing on the more promising regions of the space.

References

[1] Gat-Viks, I., Tanay, A., Raijman, D. and Shamir, R. 2005. The factor graph network model for biological systems. *Proceedings of Research in Computational Molecular Biology*, LNCS 3500, pp 31–47.

[2] Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P. and Kummer, U. 2006. COPASI—a COmplex PAthway SImulator. *Bioinformatics*, 22(24):3067–3074.

[3] Kschischang, F.R., Frey, B.J. and Loeliger, H.A. 2001. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 42(2):498–519.

[4] Koh, G., Teong, H.F.C., Clement, M.V., Hsu, D. and Thiagarajan, P.S. 2006. A decompositional approach to parameter estimation in pathway modeling. *Bioinformatics*, 22(14):e271–e280.

[5] Moles, C.G., Mendes, P. and Banga, J.R. 2003. Parameter estimation in biochemical pathways: A comparison of global optimization methods. *Genome Research*, 13(11):2467–2474.

[6] Yedidia, J.S., Freeman, W.T. and Weiss, Y. 2002. Understanding belief propagation and its generalizations. In: *Exploring Artificial Intelligence in the New Millenium*, pp 239–269.

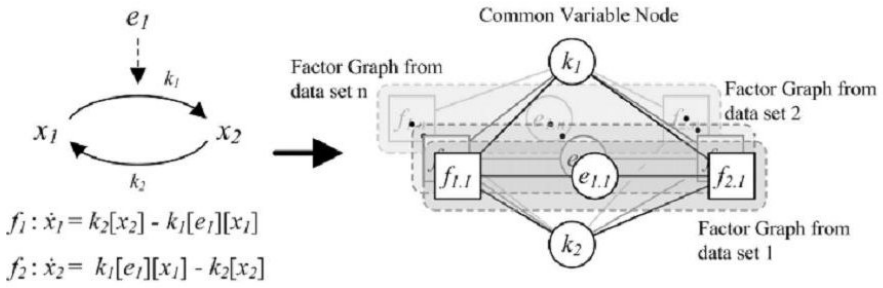


Figure 1: An enzyme mediated reaction and its factor graph representation. A factor graph is constructed for each new dataset, and they are fused at their common variable nodes (k_1 and k_2) to form a composite factor graph.

	BP	SRES	GA
1 Dataset	0.412	0.483	61.96
2 Datasets	1.548	1.356	17.38
3 Datasets	1.250	3.020	263.55
4 Datasets	0.203	2.040	46.76

Figure 2: Performance comparison of three methods on parameter estimation. BP is our method based on belief propagation. SRES and GA are two methods based on evolutionary strategies with stochastic ranking and genetic algorithms, respectively. The scores are the weighted mean squared difference between simulated and experimental data. Smaller scores are better.

Knowledge-Based Pathway Optimization Strategy for Gene Expression Profiling Analysis

Kazuo Iida,[1] Takako Takai-Igarashi,[2] Daiya Takai,[3] Hiroshi Tanaka[4]

1 Introduction

Expression profiling using DNA microarrays has become a powerful tool for discovering protein expression responses to disease. Gene Set Enrichment Analysis (GSEA) has been developed for evaluation of microarray data at the level of gene sets (for example pathways) for association with disease phenotype [1, 2]. GSEA is designed to detect subtle but coordinated changes in expression data. This approach has been successful in identifying oxidative phosphorylation as a pathogenetic pathway in diabetes [2].

JavaGSEA, a desktop application version, is freely distributed from the Broad Institute [3] and becomes popular. However, there seem no more prominent success stories in identifying novel disease related pathways. We consider that there is a certain limitation in the approach of GSEA. GSEA is designed to detect better scored pathways among pathways contained public databases. They are ideal pathways representing conceptualized molecular causalities common to similar but not identical a variety of cellular states. Subtle but certain differences among individual cellular states are omitted from them. Our improvement is based on an idea that GSEA detecting coordinated alternations in gene expression might more nicely fit with pathways not at the conceptual level but at the level of individual pathways specific to individual disease phenotypes. This improvement might enable you to obtain better results from GSEA analysis when you modify and optimize a pathway downloaded from a public database so as to be scored best on the basis of expert knowledge specific to a target disease phenotype.

Here we show a new strategy for applying GSEA to identify a new pathway contributing health risks of obstructive sleep apnea syndrome (OSAS). We took a strategy where the score goes to be optimized by iterative modification of a pathway (a hypothesis) based on expert knowledge (Figure 1).

2 Results

OSAS is a common disease characterized by recurrent collapse of the upper airway during sleep and associated with an increased health risk of hypertension, type II diabetes, angina, myocardial infarction, congestive heart failure, stroke, and fetal cardiovascular events [4]. A murine model of OSAS revealed that intermittent hypoxia (IH), namely the starvation of oxygen in a cell, could induce health risks of OSAS [4, 5]. However microarray data from the murine model did not conclude with identification of pathways contributing to the health risks of OSAS [5].

We made a second analysis of the expression profiling data of [5]. We downloaded their microarray data (ID: GSE1873) from Gene Expression Omnibus (GEO) [6]. This is a data sets examining IH effect on liver gene expression in leptin deficient (ob/ob) mice with Affymetrix GeneChip Mouse Expression Set 430 Array MOE430A.

Prior to GSEA analysis, we investigated 22 reference articles of OSAS and tried to find IHinduced responses established in other species. Fortunately we found that in fission yeast the lack of oxygen activated hypoxia induced SREBP feedback regulation of cholesterol [7]. Because increase in serum cholesterol in diabetic patients has been implemented as a major cause of OSAS-outcomes, we hypothesized that a pathway homologous to yeasts IH response was also induced in mammalian tissue by IH.

We started GSEA analysis with the homologous pathway and its associated pathways. We referred BioCyc [8] in learning homologous relations of pathways in species. According to the strategy in Figure 1, we modified the pathway several times iteratively based on expert knowledge obtained from the reference articles so as to be scored better step by step. The final best-scored pathway is shown in Figure 2 (False Discovery Rate ¡ 0.001). Although we manually optimized the pathway with try-and-error in this study,

[1]Research Institute of Bio-system Informatics, Tohoku Chemical Co.,Ltd., Japan. Email: iida@t-kagaku.co.jp
[2]Univ. Center for Information Medicine, Tokyo Medical and Dental University, Japan. Email: takai@cim.tmd.ac.jp
[3]Dept. of Clinical Laboratory, The University of Tokyo Hospital, Japan. Email: dtakai-ind@umin.ac.jp
[4]Univ. Center for Information Medicine, Tokyo Medical and Dental University, Japan. Email: tanaka@cim.tmd.ac.jp

the optimization process can be summarized in the style of algorithm, which we plan to implement in software in the future (Figure 3). In the implementation, we will be able to use protein-protein interaction databases as a data resource where extended genes are taken from.

The optimized and best-scored pathway consists of *cholesterol synthesis*, *SREBP gene regulation* and *cholesterol uptake* (Figure 2, 4). In our optimization steps, we excluded pathways of *HIF-hypoxia response* and *cholesterol secretion* because both showed no coordinated changes. The resulted best-scored pathway enables us to interpret a molecular mechanism causing health risks of OSAS induced by IH (Figure 4): 1) Because *cholesterol synthesis* pathway consumes too much oxygen to run under the starvation of cellular oxygen [7], there is a drop in cholesterol in a cell. 2) SREBP detects the drop of cholesterol and promotes transcription of genes required in cholesterol synthesis. 3) SREBP also up-regulates LDLR so as to increase cellular cholesterol by uptaking cholesterol from serum. Our identification of this pathway activated in OSAS-model mice shows a piece of evidence that *hypoxia induced SREBP feedback regulation of cholesterol*, established in fission yeast, may also exist in mammalian cells (Table 1). This pathway might be evolutionally conserved in eukaryotic kingdom, from fission yeast to mammalian organisms.

In this study we could extract new knowledge by a second analysis of microarray data in GEO. GEO keeps increasing in its entries. We believe that you can identify much more disease-related pathways using GSEA analysis with our knowledge-based pathway optimization strategy.

References

[1] Subramanian A, et al. (2005) *Proc Natl Acad Sci USA*, 102:15545.

[2] Mootha VK, et al. (2003) *Nat Genet*, 34:267.

[3] `http://www.broad.mit.edu/gsea/`.

[4] Polotsky VY, O'Donnell CP (2007) *Proc Am Thorac Soc*, 4:121.

[5] Li J, et al. (2005) *J Appl Physiol*, 99:1643.

[6] `http://www.ncbi.nlm.nih.gov/geo/`.

[7] Hughes AL, et al. (2007) *J Biol Chem*, 282:24388.

[8] `http://www.biocyc.org/`.

[9] Hughes AL, et al. (2005) *Cell*, 120:831.

Abbreviations used in text: SREBP (sterol regulatory element binding protein), LDLR (low-density lipoprotein receptor).

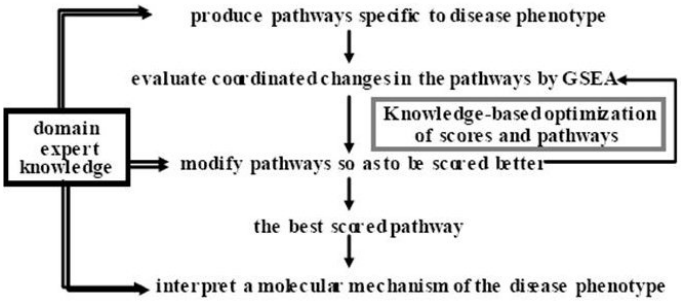


Figure 1: Our strategy to find disease-related pathways from DNA microarray data.

Species	Hypoxia induced SREBP feedback regulation of cholesterol	Evidence
Fission yeast	Yes	Huges, AL (2005) [9]
Mammalian	Yes	This study

Table 1: An evolutionally conserved hypoxia response.

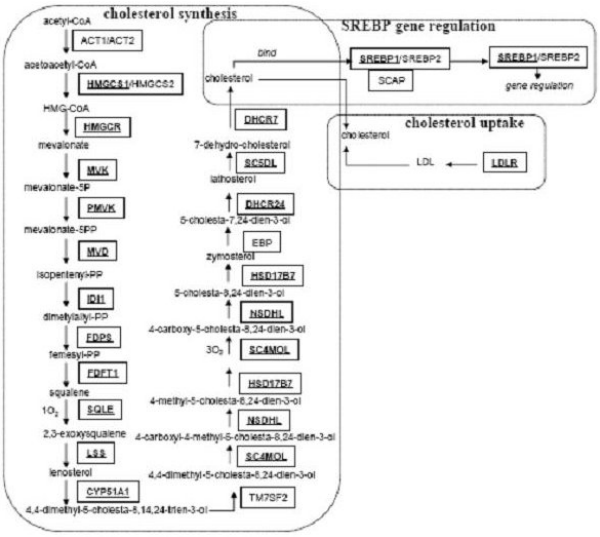


Figure 2: The best-scored pathway in GSEA analysis with knowledge-based optimization. Genes with underline showed significant coordinate changes in the expression profile. Genes without underline showed no coordinated changes.

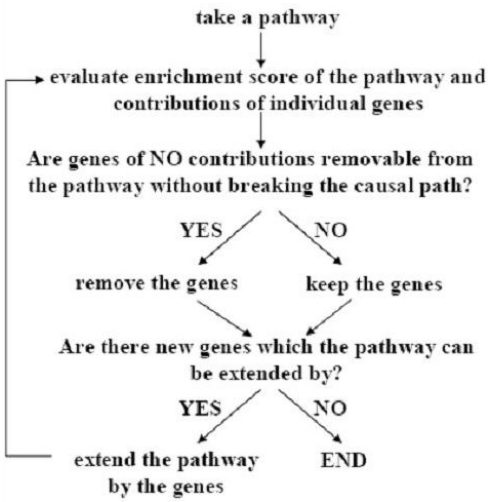


Figure 3: Pathway optimization algorithm.

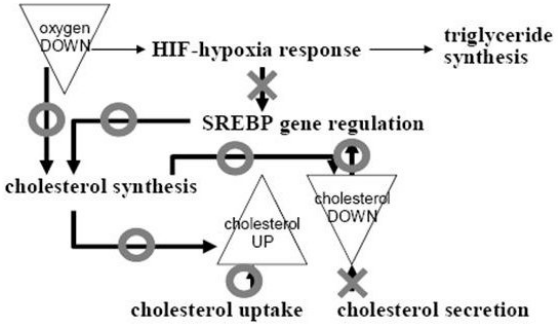


Figure 4: Hypoxia induced SREBP feedback regulation of cholesterol in a mammalian cell suggested by this study.

Reduced CpG Mutation Rate Suggests Functional Role of Intragenic and 3 CpG Islands in Human Genes

Julia Medvedeva,[1] Marina Fridman,[2] Nina Oparina,[3] Dmitri Malko,[4] Ekaterina Ermankova,[5] Ivan Kulakovsky,[6] Vsevolod Makeev[7]

CpG islands (CGIs) are usually defined as DNA segments that are longer than 200 bp, have over 50% of G+C content, and have CpG frequency of at least 0.6 of that statistically expected [2]. Most of the studies focused on CpG islands considered CGIs associated with 5' gene regions. Generally, such islands are more than 1 kb long, can cover the promoter, TSS, the first coding exon and have stronger parameters of G+C content and Obs/Exp [3]. The methylation status of such CGIs is believed to influence the transcription level of a corresponding gene.

Contrary to the widespread opinion mentioned above only 50% of CGIs are located near TSS. About 20% gene-associated CGIs are disposed in internal and 3' terminal gene regions. Internal exons display less overlapping with CGIs than exons in 5' regions and to some extend exons in 3' regions of the genes. CGIs associated with 3' region of the gene are more often overlapped with coding exons than with 3' UTR [2].

So far the question arises if CGIs observed in protein-coding regions can be considered as a result of protein selection. We decided to evaluate selection at genome and protein levels and mutation rate at CpG sites belonging to 5'-assosiated, intragenic and 3'-assosiated CGIs. To this end we compared mutation rate and selection in exons overlapping and not overlapping with CGIs separately for non-CpG containing codons (the background) and CpG containing codons.

Thus, we calculated dn/ds ratio [1] in human-mouse alignments. We also used dn and ds values separately. The results are presented in Table 1.

	1st exon			internal exon			last exon		
Codon type	dn	ds	dn/ds	dn	ds	dn/ds	dn	ds	dn/ds
CG pair in CGI	0,131	0,512	0,257	0,097	0,910	0,106	0,100	0,800	0,125
CG pair out of CGI	0,136	0,987	0,138	0,093	1,510	0,061	0,114	1,273	0,090
AG pair in CGI	0,146	0,485	0,302	0,101	0,644	0,157	0,109	0,599	0,181
AG pair out of CGI	0,134	0,508	0,264	0,087	0,535	0,164	0,112	0,533	0,210
GC pair in CGI	0,130	0,381	0,342	0,098	0,534	0,183	0,101	0,503	0,201
GC pair out of CGI	0,145	0,488	0,297	0,095	0,526	0,180	0,122	0,519	0,235
GA pair in CGI	0,120	0,381	0,314	0,084	0,531	0,159	0,091	0,463	0,197
GA pair out of CGI	0,114	0,450	0,252	0,075	0,489	0,154	0,096	0,479	0,200

Table 1: Dn/Ds test and dn, ds-values separately calculated for different types of codon.

Conclusions:

1. CpG island decrease mutation rate in CpG pairs at synonymous sites approximately twofold.
2. Effect of CGI does not depend on the exon location within the gene: 5'-assosiated, intragenic and 3'-assosiated CGIs protect CpG sites from methylation and probably play the same regulatory role in gene functioning.

[1]Institute of Genetics and Selection of Industrial Microorganisms, Russia. Email: ju.medvedeva@gmail.com
[2]Institute of Genetics and Selection of Industrial Microorganisms, Russia. Email: marina-free@mail.ru
[3]Engelhardt Institute of Molecular Biology, RAS, Russia. Email: oparina@gmail.com
[4]Institute of Genetics and Selection of Industrial Microorganisms, Russia. Email: carbonoid@mail.ru
[5]Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Russia. Email: ermakova8@yandex.ru
[6]Institute of Genetics and Selection of Industrial Microorganisms, Russia. Email: ikulakovsky@inbox.ru
[7]Institute of Genetics and Selection of Industrial Microorganisms, Russia. Email: makeev@genetika.ru

Auxiliary conclusions:

1. CGIs located in 5' gene regions better protect CpG pairs from mutations. The substitution rate in synonymous CpG site is lowest for islands located in 5' gene segments.

2. CGIs caused a very weak selection at DNA level. ds, dn calculated for nonCpG codons in CpG island and out of CpG islands are practically identical.

3. Protein selection is strong enough to overcome effect of CGIs. In non-synonymous CpG sites the substitution rate is practically identical for such sites covered and non-covered with CGIs. Thus, the majority of mutations become eliminated with selection.

4. Proteins are more variable at their ends: the number of non-synonymous substitutions is greater in terminal exons than in internal exons.

5. Protein selection at 3' region of the gene is weaker. The substitution rate in non-synonymous CpG sites within islands is less than that out of islands.

6. CpG pairs are nevertheless under selection at DNA level. The number of synonymous substitutions in CGIs is the highest near the center of gene, than in 3' region of the gene and is the least in the 5' region of the gene, which is probably associated to the protein factor binding sites near the gene terminals related to transcription initiation and termination (or anitsense transcription initiation, microRNA binding etc).

References

[1] Ina, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J. Mol. Evol.*, 40:190–226.

[2] Gardiner-Garden, M. and Frommer, M. 1987. CpG islands in vertebrate genomes. *J. Mol.Biol.*, 196:261–282.

[3] Ponger L., Duret L., Mouchiroud D. 2001. Determinants of CpG Islands: Expression in Early Embryo and Isochore Structure. *Genome Research*, 11:1854–1860.

Structural Modeling of MaSp1 and MaSp2 Proteins of Dragline Silk in *Latrodectus hesperus*

Saboura Ashkevarian,[1] Armin Madadkar Sobhani,[2] Bahram Goliaei[3]

1 Introduction

Dragline silk with extraordinary mechanical properties, is one of the toughest materials known in the world [3]. Being a biocompatible and biodegradable material [5], the dragline silk is an interesting substance for commercial purposes and medical researches.

Despite being in the focus of biologists and material scientists for centuries, man could not produce fibers with the same toughness as dragline silk or even propose an approved model for this protein. Therefore, according to accessible theoretical methods, in this research we proposed a suitable model for dragline silk protein using molecular modeling and computational biology techniques and software.

It has been suggested that the dragline silk is consisted of two large proteins: major ampullate spidroins (MaSp1) and MaSp2 [4]. Recently, the whole sequence of these two proteins were resolved in full length for Black Widow spider (*Latrodectus hesperus*) by Hayashi et al. with amino acid lengths of 3129 and 3779, respectively [2]. Since these proteins are made of alternating polyalanine and glycine-rich blocks between non-repetitive N- and C- terminal domains, we modeled one of the most common repetitive parts known as consensus sequence with length of 35 and 24 amino acids for MaSp1 and MaSp2 proteins of *L. hesperus*.

2 Methodolog

Using dihedral angles of GGA motifs reported previously [1], we constructed 3D structure of GGAGQGGA sequence of MaSp1 by HyperChem 7.5 and carried out an extensive conformational search for four GQ dihedral angles. Then resulting structure was elongated to 35 consensus sequence and used as an input for a 5 ns molecular dynamics run using GROMACS 3.3.2 package. Also using a template (PDB code:1DAB) and MODELLER 9v2 software, we modeled MaSp2 consensus of 24 amino acids. Final structures were analyzed using PROCHECK and GROMACS analysis tools

3 Results and Discussion

Our results approves that this fibers are less ordered in water. But they are in helical and beta sheet conformations when they are in no water condition. It seems that there is no helical or beta sheet structures in the abdomen of spider but when they are in their solid form out of spiders body they are ordered in helical and beta sheet conformations. Our models and their analysis are shown in Figure 1.

References

[1] Ashida J, Ohgo K, Komatsu K, Kubota A, Asakura T. Determination of the torsion angles of alanine and glycine residues of model compounds of spider silk (AGG)10 using solid-state NMR methods. *Journal of Biomolecular NMR*, 25(2):91-103, 2003.

[2] Ayoub NA, Garb JE, Tinghitella RM, Collin MA, Hayashi CY. Blueprint for a high-performance biomaterial: Full-length spider dragline silk genes. *PLoS ONE*, 2:e514, 2007.

[3] Gosline JM, Guerette PA, Ortlepp CS, Savage KN. The mechanical design of spider silks: From fibroin sequence to mechanical function. *Journal of Experimental Biology*, 202(23):3295-3303, 1999.

[1]Laboratory of Biophysics and Molecular Biology, Institute of Biochemistry and Biophysics, University of Tehran, Iran. Email: s.ashkevarian@ibb.ut.ac.ir

[2]Department of Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Iran. Email: armin@ibb.ut.ac.ir

[3]Department of Biophysics and Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Iran. Email: goliaei@ibb.ut.ac.ir

[4] Hinman MB, Lewis RV. Isolation of a clone encoding a second dragline silk fibroin. Nephila clavipes dragline silk is a two-protein fiber. *Journal of Biological Chemistry*, 267(27):19320–19324, 1992.

[5] Kaplan DL. Fibrous proteins—silk as a model system. *Polymer Degradation and Stability*, 59(1-3):25–32, 1998.

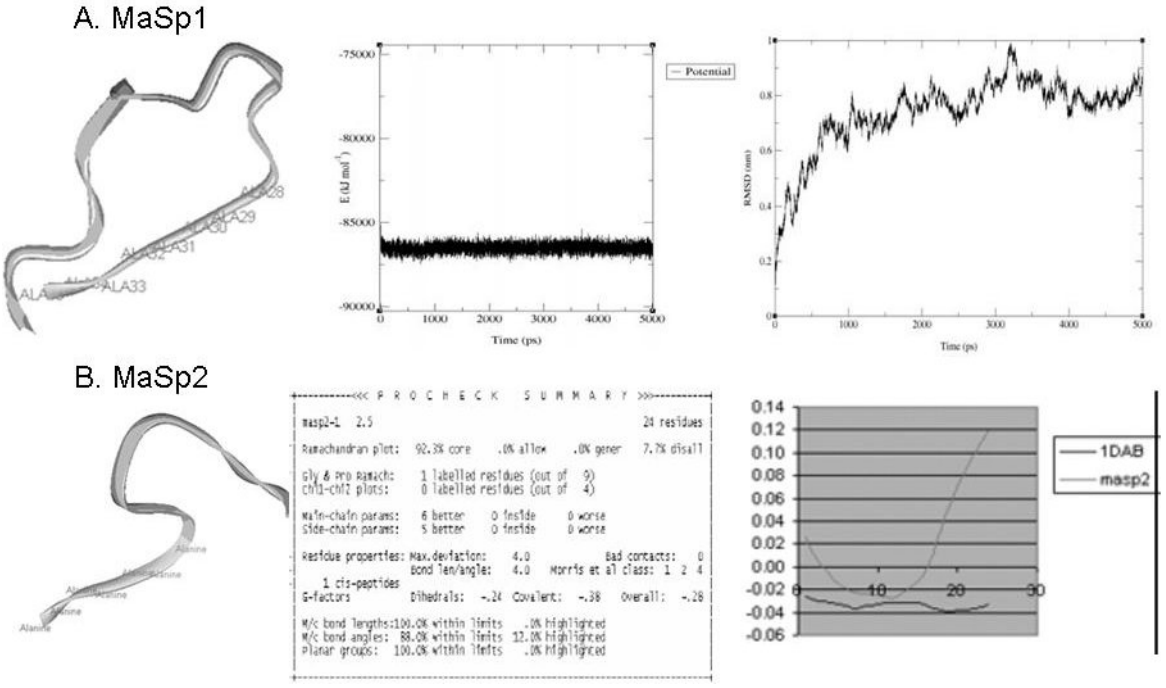


Figure 1: A and B structure of both proteins and their analysis.

Missing Values Estimation for DNA Microarray Gene Expression Data: SPLS[1]

Kyungsook Kim,[2] Mira Oh,[3] Jangsun Baek,[4] Young Sook Son[5]

1 Introduction

DNA Microarray gene expression data often contain missing values because their size is very big and also their observation process is very complex. Most of statistical methods are applied after genes with missing values are excluded even though there is only one missing value. For the efficient use of data with a few missing values, it is desirable that missing values are replaced with their accurate estimates.

The initial simple approaches to deal with missing values are such as deleting genes with missing values, imputing missing values to zero, or imputing missing values of a certain gene to the average of the values observed in the gene. After that, as the approaches using the local structure of data were introduced K-nearest neighbor imputation (KNN) by Troyanskaya et al. (2001), sequential KNN imputation (SKNN) by Kim et al. (2004), and local least squares imputation(LLS) by Kim et al. (2005), etc. On the other hand, as the approaches using the global structure of data were introduced Bayesian principal component analysis imputation (BPCA) by Oba et al. (2003) and partial least squares regression imputation (PLS) by Nguyen et al. (2004), etc.

We propose sequential partial least squares regression imputation (SPLS) to estimate missing values for time-course gene expression data that have correlations among observations over time points.

2 Sequential Partial Least Squares Regression Imputation

SPLS, the way using locally data, uses selectively K genes most similar to the target gene with missing values among all complete genes without missing values. The target gene is sequentially determined as ascending order in the number of missing values.

When all missing values in the target gene are estimated by PLS regression, it is turned into a new complete gene. Then this gene becomes a candidate in selecting K genes for the next target gene. SPLS-gene method is performed sequentially by PLS regression in gene-wise way, and also SPLS-array method is performed in array-wise way. Additionally, SPLS-combined method is carried out by combining these two ways.

3 Results of Simulation

Three yeast data, all time-course data, are applied to compare the performance of SPLS and some former methods. Y7 data in DeRisi et al. (1997) have 7 time points. Y18 and Y24 data in Spellman et al. (1998) have 18 time points from alpha-factor part and 24 time points from CDC15 part, respectively.

Simulated data of three types are generated from each yeast data. Type 1 keeps the intrinsic missing structure of original data. Type 2 assigns artificially several missing rates under 20%. Type 3 is intended to exclude the dependence on only one data simulated as Type 2. So, we make Type 3 data by repeating 100 times Type 2 experiment. The performance of the missing value estimation is evaluated by normalized root mean square error (NRMSE). Here is presented only the result of Y18 data because of space limited. For SPLS the best result is selected among results by SPLS-gene, SPLS-array, and SPLS-combined. As results, SPLS is superior to any other methods compared as shown in the figure and the table.

[1]This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD)(KRF-2005-204-C00017).

[2]Department of Statistics, Chonnam National University, Korea. Email: ksook620@jnu.ac.kr
[3]Department of Statistics, Chonnam National University, Korea. Email: omr@chonnam.ac.kr
[4]Department of Statistics, Chonnam National University, Korea. Email: jbaek@chonnam.ac.kr
[5]Department of Statistics, Chonnam National University, Korea. Email: ysson@chonnam.ac.kr

References

[1] DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686.

[2] Kim, H., Golub, G.H. and Park, H. (2005). Missing value estimation for DNA microarray gene expression data: Local least squares imputation. *Bioinformatics*, 21:187–198.

[3] Kim, K.Y., Kim, B.J. and Yi, G.S. (2004). Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinformatics*, 5:160.

[4] Nguyen, D., Wang, N. and Carroll, R. J. (2004). Missing value estimation for cancer microarray gene expression data. *Journal of Data Science*, 2:347–370.

[5] Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K. and Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19:2088–2096.

[6] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Bostein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Ssccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9:3273–3297.

[7] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Bostein, D. and Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520–525.

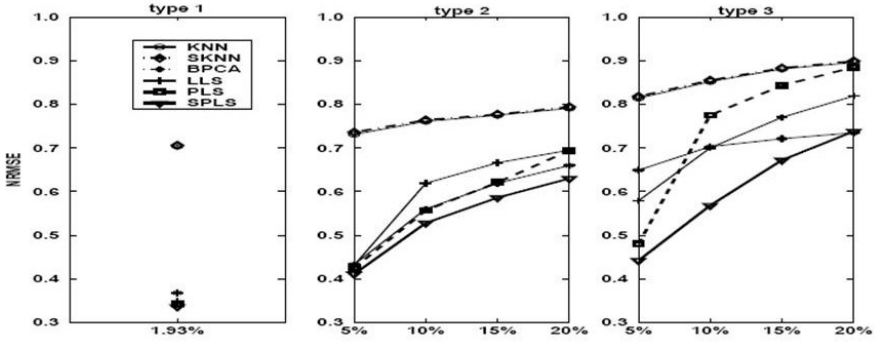


Figure 1: Comparison of the NRMSEs for Y18 dataset.

Experiment Type	Missing Rate(%)	KNN	SKNN	BPCA	LLS	PLS	SPLS
Type 1	1.93	.704	.706	.368	.348	.342 (Ar, C)	.336 (Ar, D)
Type 2	5	.731	.736	.433	.431	.425 (Ar, D)	.411 (Ar, D)
	10	.761	.764	.560	.619	.557 (Ar, AC)	.528 (Ar, D)
	15	.775	.777	.619	.666	.621 (Ar, AC)	.586 (Ar, AC)
	20	.791	.794	.660	.695	.694 (Ar, AC)	.630 (Ar, AC)
Type 3	5	.814(.03)	.818(.03)	.649(.05)	.579(.05)	.480(.11) (Ar, AC)	.442(.05) (Ar, AC)
	10	.852(.02)	.855(.02)	.703(.08)	.700(.04)	.775(.03) (Gn, AC)	.568(.04) (Ar, AC)
	15	.881(.02)	.883(.02)	.721(.03)	.770(.03)	.843(.03) (Gn, AC)	.672(.04) (Ar, AC)
	20	.896(.02)	.899(.02)	.735(.03)	.819(.03)	.885(.03) (Gn, AC)	.738(.03) (Ar, AC)

* The numbers in the cells of Type 3 imply the mean(the standard deviation) of NRMSEs on 100 replications.
* (a, b) in the PLS or SPLS of Type 3 denotes the best case among a: local ways or b: similarity measures.
* Ar: array-wise way, Gn: gene-wise way.
* C: correlation, D: Euclidean distance, AC: absolute correlation.

Table 1: Comparison of the NRMSEs for Y18 dataset.

Analysis of Human Cis-Antisense Transcription: Primate-Specific Exonic Sequences, Structure-Dependent Sense-Antisense Co-Expression, and functionally Restricted Noncoding-RNA Transcription

Leonard Lipovich,[1,2] Charlie W-H. Lee,[3] Hui Jia,[3] Yuri Orlov,[3] Thomas Wee-Hong Ng,[3] Jieming Chen,[4] Edwin Lian-Chong Ng,[5] Edison T. Liu,[3] Lance D. Miller,[3] Lawrence W. Stanton,[3] Ken W.-K. Sung,[6] Vladimir A. Kuznetsov[2,7]

1 Introduction

A cis-antisense gene pair contains two distinct genes mapping to opposite strands of the same locus and transcribed in opposite orientations with partially overlapping exonic sequence. Such genes may regulate each other, pre- or post-transcriptionally. Thousands of cis-antisense gene pairs, which generally contain long protein-coding genes and/or mRNA-like noncoding-RNA (ncRNA) genes, rather than hundreds of short- RNA precursors, reside endogenously in eukaryotic genomes. The shared exonic sequence in each pair is defined as the exon-to-exon cis-antisense overlap (Figure 1), distinct from the remaining, nonoverlapping exonic sequences of the paired genes Cis-antisense is abundant and occurs in all kingdoms of life. Up to 25% of mammalian genes reside in cis-antisense pairs [1, 2]. Conservative estimates of the total number of human cis-antisense pairs ranged from hundreds to approximately 2500. With increasing sophistication of computational antisense discovery tools and ongoing growth of cDNA and EST databases [2], estimates increased to 4000-6000 [3, 4]. Many examples of natural cis-antisense ncRNAs overlapped with either protein-coding or non-coding genes have been recently described. However, expression, functions and evolution origin of these common and other more specific classes of cis-antisense transcripts have not been systematically studied. In addition, genomewide analyses of mammalian cis-antisense are generally based on automated pipelines, which are vulnerable to artefacts in transcriptome databases.

To elucidate expression patterns, functions and evolution of human cis-antisense, we constructed a genomewide, nonredundant cis-antisense catalog by careful manual curation, identifying 4,511 cis-antisense pairs. 52% of cis-antisense overlaps were inside protein-coding open reading frames.

2 Results

We addressed these problems by integrating the output of several published human cis-antisense discovery efforts with the results of our own cis-antisense pipeline, and by then manually curating each non-redundant cis-antisense locus to eliminate the artifacts. We then systematically analyzed genomic structure (head-tohead, tail-to-tail, or embedded), protein-coding capacity, and gene functions of all cis-antisense pairs. We combined cDNA/EST, longSAGE, and microarray data to analyze sense and antisense co-expression.

[1]Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI 48201. Email: LLIPOVICH@MED.WAYNE.EDU
[2]Joint corresponding author.
[3]School of Chemical and Life Sciences, Singapore Polytechnic, Sinngapore 139651
[4]Department of Biological Sciences, National University of Singapore, Singapore 117543
[5]School of Chemical and Life Sciences, Nanyang Polytechnic, Singapore 569830
[6]Genome Institute of Singapore, Singapore 138672.
[7]Bioinformatics Institute, Singapore 138671. Email: VLADIMIRK@BII.A-STAR.EDU.SG

We found that noncoding-RNA (ncRNA) cis-antisense transcription is significantly associated with highly specific functional gene categories, including developmental, Wnt signaling, and neuronal genes. Surprisingly, 803 noncoding-RNA members of cis-antisense pairs harbored primate-specific Alu repeats in their exons, although exonic Alu incidence in cis-antisense noncoding-RNA genes was depleted relative to genomic background. 554 of those genes had transcription starts, splice sites, and/or 3' ends within Alu repeats, with enrichment for the oldest (AluJ) subclass. Therefore, hundreds of cis-antisense pairs have been likely originated early in primate evolution. Finally, we mapped U133A & B chip target sequences on gene loci of cis-antisense gene pairs and found that ∼20% of all U133A & B target sequences mapping such genes. In different tissues (including normal and cancerous brain, breast, lung) the concordant (co-regulation) expression pattern was dominated versus anti-regulation pattern. This result is consisted with our previous observation reported for human breast cancer tissues [4]. We also found that sense-antisense co-expression depend on genomic organization of the gene partners, with embedded pairs co-expressed less often than convergent or divergent pairs.

3 Conclusion

The importance of our discoveries is in the revelation that cis-antisense regulation is predominantly associated uniquely with noncoding-RNA transcription at cis-antisense loci. A second important result is that Alu insertions in primate evolution, however, have not played a distinct functional role in noncoding-RNA cis-antisense regulation. Furthermore, we are the first group to uniquely relate sense-antisense co-expression to the genomic structure type of cis-antisense pairs.

In-depth stratification of physically overlapping anti-parallel genes by genomic architecture, expression analysis and evolutionary history therefore reveals additional dimensions in the functional space of the human transcriptome, and biological insights absent from earlier antisense studies.

Our approach demonstrates that careful analysis of the genomic structure and architecture of cis-antisense loci, when combined with expression data, is capable of unraveling functional signals previously unseen in cis-antisense data, and therefore should facilitate future work.

References

[1] Chen, J. et al. 2004. Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res*, 32:4812–4820.

[2] Zhang, Y. et al. 2006. Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Res*, 34:3465–3475.

[3] Engstrom, P. G. et al. 2006. Complex loci in mammalian genomes. *PLoS Genetics*, 2:e47.

[4] Kuznetsov, V. A. et al. 2006. Genome-wide co-expression patterns of human cis-antisense gene pairs. In: *Proc. 5th International Conf. on Bioinformatics for Genome Regulation and Structure*, Institute of Cytology and Genetics, Novosibirsk, Russia. v.1: 90–93.

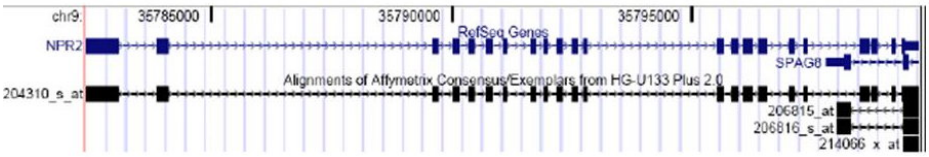


Figure 1: Overlapping protein-coding transcripts of NPR2-SPAG8 pair. The variants of transcripts are overlapped by corresponding exons. NPR2 is encoding natriuretic peptide receptor B/guanylate cyclase B (atrionatriuretic peptide receptor B). According AceView program, at least 7 spliced variants could be observed. This gene is expressed at moderate or high levels in many tissues. SPAG8 is sperm associated antigen 8 isoform 2. According AceView program, at least 11 spliced variant of this genes could be observed. This gene is moderately expressed in more than 40 tissues. Affymetrix U133 probesets strongly support the expression of these genes.

HIV-1 Protease Inhibitor Comparative Docking Studies of Synthetic and Natural Compounds

Rajeev Kumar,[1] Manmath Routray,[2] J. Febin Prabhu Dass[3]

1 Introduction

Nature has always provided a source of drugs for various ailments. A number of medicinal plants have been reported to have anti-HIV properties. The bioactivity-guided fractionation of crude extracts has provided lead molecules for discovery of anti-HIV drug candidates. A variety of secondary metabolites obtained from natural origin showed moderate to good anti-HIV activity.

2 Method and Results

We took five natural compounds Nomilin and Limonin [1] isolated from *citrus spp.* (Family Rutaceae), Uvaol and ursolic [2] acid isolated from the methanolic extract of leaves of *Crataegus pinatifida* (family Rosaceae), and Maslinic acid [3] isolated from *Geum japonicum.* These showed potent inhibitory activity against HIV-1 protease. We docked each compound with protease and obtained the result, then we compare the binding site of these molecule with existing drug molecule.

Synthetic Compound	Score	Area	ACE	Transformation
Amprenavir	9822	1243.30	-750.08	0.20 0.09 -0.01 1.32 2.06 0.54
Tripnavir	8822	981.10	-280.15	1.79 -0.03 -1.44 24.67 14.73 -2.33
Nefilavir	7816	886.60	-374.36	1.41 -0.02 3.14 10.77 -14.26 12.37

Natural Compound	Score	Area	ACE	Transformation
Nomilin	6272	711.10	-253.22	2.34 0.67 -0.84 6.05 -4.83 16.67
Uvaol	6234	692.80	-273.73	0.27 0.88 -2.41 19.61 -0.59 5.05
Maslinic acid	6198	679.20	-261.30	-0.38 -0.47 0.93 -7.21 -9.88 18.81
Ursolic acid	6174	677.30	-266.29	0.18 0.80 -2.32 20.04 1.10 5.32
Limonin	5610	600.10	-246.14	-2.63 -0.49 2.24 4.49 -1.71 14.49

3 Conclusion and Future Work

After getting the docking results we can see that there is similarity between synthetic and natural compounds. The overall docking score is less than original existing drug but after increasing the bulkiness of natural compounds or lead optimization the score may increase than existing drug. On visualizing the receptor-ligand complex of both synthetic and natural compounds, It observed that both having similar binding pocket. Although no plant-derived drug is currently in clinical use to treat AIDS, promising activities shown by these natural compounds can be taken into further account. For future work we are working on QSAR studies of these protease inhibitor natural products. That will give more clear picture of these natural products.

References

[1] Battinelli, L., Mengoni, F., Lichtner, M., Mazzanti, G., Saija, A., Mastroianni, C. M. and Vullo, V. Effect of limonin and nomilin on HIV-1 replication on infected human mononuclear cells. *Planta Med.*, 2003, 69:910–913.

[2] Min, B. S., Jung, H. J. and Lee, J. S. Inhibitory effect of triterpenes from Crataegus pinatifida on HIV-1 protease. *Planta Med.*, 1999, 65:374–375.

[3] Xu, H.-X., Zeng, F.-Q., Wan, M. and Sim, K.-Y. Anti-HIV triterpene acids from Geum japonicum. *J. Nat. Prod.*, 1996, 59:643–645.

[4] *Nature review*, December 2007, Vol. 6, No. 12.

[1]VIT University, Vellore, India. Email: rajeev.vit@gmail.com
[2]VIT University, Vellore, India. Email: manmathroutray@gmail.com
[3]VIT University, Vellore, India. Email: mail2febin@gmail.com

A DNA-Based Algorithm for Calculating the Maxflow in Networks

Andrea Sackmann,[1] Piotr Formanowicz[1,2], Jacek Błażewicz[1,2]

1 Introduction

DNA Computing was introduced in 1994 when Leonard M. Adleman showed that DNA, working in nature as a kind of Turing Machine, can be used for solving computational problems. He encoded a directed graph in the base sequences of DNA molecules and found a Hamiltonian path through the graph by manipulation the molecules by means of molecular biology operations performed in a wet laboratory [1]. Several months later Richard J. Lipton generalized Adleman's approach by proposing a DNA based solution of the satisfiability (SAT) problem [4]. The subsequent works mainly considered the DNA Computing from a theoretical viewpoint concerning with automata theory, formal language theory or computability models, as e.g. defined by so-called sticker systems [5]. Put into practice, several mathematical decision or combinatorial optimization problems have been solved by means of DNA. The theoretically outperforming advantages inherent in the strategy of using DNA molecules for computation are the extreme density of information storage (one bit can be stored in about one cubic nanometer) as well as the massive parallelism of the approach. One test tube can hold up to 10^{20} strands of DNA and a molecular biology technique is performed on each of those strands in parallel. Furthermore, DNA Computing is very energy efficient since approx. 2×10^{19} operations are performed per joule.

With this work we introduce a DNA based algorithm calculating the maximum flow of a network. The maxflow problem, formally defined e.g. in [2], is a classical optimization problem with many applications in different fields and corresponding silicon based algorithms have been studied for over four decades, cf. [3]. These conventional algorithms are polynomial in time. To the best of our knowledge this is the first application of DNA Computing in the field of flow networks and our main goal is to demonstrate the possibility of solving problems of this kind by DNA molecules.

2 The calculation

We introduce an algorithm calculating the maxflow of a network by solving its dual, i.e. by means of finding the capacity of a minimal cut of the network.

First, as input to the algorithm several molecules are synthesized encoding information units (YES) or (NO) for each vertex and each edge as well as the edge weights. Adding of molecules being Watson-Crick complement half of one and half of another information unit leads (as a kind of bridges) to a concatenation of the information units to so-called network templates. After the removing of the bridge molecules, the templates are single stranded. They constitute the combinatorial library of the approach and encode all possible combinations of information units of each vertex and each edge (where they include exactly one unit per vertex and exactly one per edge). A positive information unit (YES) of a vertex means that this vertex is an element of the set S in which the source of the network is included while a negative information unit (NO) stands for its belonging to the set T which also contains the sink. This already gives a partition of the vertex set. A positive information unit (YES) of an edge indicates that this edge is directed from set S to set T while a negative information unit (NO) implies that the corresponding edge lays either within set S or within set T. Thus, the sum of the capacities of all edges with positive information unit (YES) is equal to the capacity of the corresponding cut. Table 1 illustrates which combinations of the information units encode a solution. Obviously, not all molecules a priori contained in the library fulfill the requirements of these combinations of information. Therefore, all units contain subsequences in which the incidence of edges to vertices is encoded. According to [6] all molecules not representing a solution for the problem contain reverse Watson-Crick complementary subsequences. Therefore, each of them fold to anneal with itself building a hairpin structure. Since the

[1]Institute of Computing Science, Poznań University of Technology, Poznań, Poland.
[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland. Email: {asackmann,piotr,jblazewicz}@cs.put.poznan.pl

start vertex	edge	end vertex	solution?
YES	YES	YES	-
YES	YES	NO	+
YES	NO	YES	+
YES	NO	NO	-
NO	NO∗	YES	+
NO	NO∗	NO	+

Table 1: All possible combinations of information contained in the information units of an edge, its start- and end vertex. The two left columns are given through template building (the information combination No-YES of a vertex and its outgoing edge is not built). The sequences of units NO∗ differ from these encoding the information NO. Dependent on the information unit on the edge's end vertex, listed in the following column, the molecule represents a solution (+) or does not (-), indicated in the right column.

double stranded stems of those hairpins contain the recognition site of a restriction enzyme, the molecules encoding no solution can be excluded from the library by adding this endonuclease to cut them (e.g. a positive vertex information unit forms a double stranded stem with the positive unit of its incoming edge). Hereafter, the uncut molecules encode cuts of the network whose capacity is given in the positive edge units. Those molecules are exponentially amplified by performing a polymerase chain reaction (PCR) with the suited primers. After separating the molecules by length by means of a gel electrophoresis the smallest molecules, which are bigger than a network constant (given by the length of several information units and the number of arcs), represent a cut with a minimal capacity. Sequencing these from the gel extracted molecules reads out the sequence which decoded give the cut itself and its capacity.

3 Results and Conclusions

We present a DNA-based algorithm to solve the maxflow problem of a given network. This algorithm is constant in time and implemented in the wet laboratory. Therefore, we show that flow network problems can be efficiently solved by means of DNA Computing.

Acknowledgements. This research has been partially supported by the EU grant Bioptrain and by the Polish Ministry of Science and Higher Education.

References

[1] Adleman, L. 1994. Molecular computation of solutions to combinatorial problems. *Science* 266:1021–1024.

[2] Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C. 2001. *Introduction to Algorithms, 2nd Ed.* Cambridge, MA: MIT Press.

[3] Goldberg, A. V. 1998. Recent developments in maximum flow algorithms. In: *Proceedings of 6th Scandinavian Workshop on Algorithm Theory* (SWAT98), pages 1–10.

[4] Lipton, R. J. 1994. DNA Solution of Hard Computational Problems. *Science* 268:524–548.

[5] Păun, G., Rozenberg, G. and Salomaa, A. 1998. *DNA Computing: new computing paradigms.* Berlin: Springer Verlag.

[6] Sakamoto, K., Gouzu, H., Komiya, K., Kiga, D., Yokoyama, S., Yokomori, T. and Hagiya, M. 2000. Molecular Computation by DNA Hairpin Formation. *Science* 288:1223–1226.

Neural Network Imputation for Missing Values in Time-Course Gene Expression Data[1]

Mira Oh,[2] Kyungsook Kim,[3] Young Sook Son[4]

1 Introduction

Missing values are often occurred due to various factors when dealing with microarray gene expression data. However, most statistical methods can not be applied to missing data. Genes with missing values must be deleted or imputed. Several imputation methods have been proposed. K-nearest neighbor imputation(KNN) by Troyanskaya et al. (2001), local least squares imputation(LLS) by Kim et al. (2005), and Bayesian principal component analysis imputation(BPCA) by Oba et al. (2003) often appear in the study on missing value estimation of microarray gene expression data.

In this study, we use artificial neural network(NN) to estimate missing values for time-course gene expression data with correlations over time points. The NN imputation is compared with KNN, LLS, and BPCA through a numerical study applied to two yeast data.

2 Neural Network Imputation

Neural network is a nonlinear and nonparametric modeling technique to solve the prediction problem in data analysis with complexity. The most widely used NN model is multiplayer perceptron(MLP). The MLP (Figure 1) consists of an input layer with input neurons connected to a hidden layer or more hidden layers with hidden neurons, which are connected to an output layer.

Missing values of the target gene in such a method as LLS are estimated by linear model. However, timecourse gene expression data can have the nonlinear relationship between missing values and observed values. This is a motive to try NN imputation. In the application of NN, missing values of target genes are put in an output layer and observed values put in an input layer. Also target genes with missing value are used as a test data set and complete genes without missing values are used as a training data set.

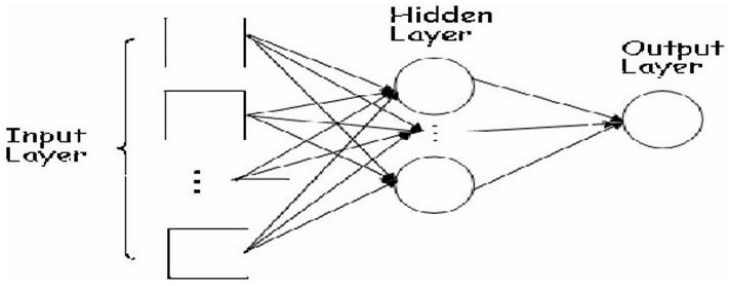


Figure 1: Architecture of feed forward neural network (MLP).

3 Numerical Results

We applied NN imputation to two yeast data, Y7 and Y24. Y7 data in DeRisi et al. (1997) have 7 time points and Y24 data from CDC15 part in Spellman et al. (1998) have 24 time points. The MLP model

[1]This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2005-204-C00017).

[2]Department of Statistics, Chonnam National University, Korea. Email: omr@chonnam.ac.kr
[3]Department of Statistics, Chonnam National University, Korea. Email: ksook620@jnu.ac.kr
[4]Department of Statistics, Chonnam National University, Korea. Email: ysson@chonnam.ac.kr

(Figure 1 used has a hidden layer. The number of hidden neurons is determined as a number giving the minimum NRMSE (normalized root mean square error) by simulation among $2 \sim 7$ hidden neurons. We used the backpropagation algorithm of the Matlab Neural Network Toolbox program. The performance of the missing value estimations is evaluated by NRMSE Table 1 shows results of NRMSEs based on KNN, LLS, BPCA, and NN imputation. NN is competitive for cases with 5% and 10% missing rates in Y7 and Y24 data.

Datasets	Missing rates	KNN	BPCA	LLS	NN
Y7	0.5%	0.560	0.567	0.533	0.533
	1%	0.606	0.615	0.597	0.608
	5%	0.583	0.596	0.568	0.487
	10%	0.630	0.644	0.618	0.609
Y24	5%	0.627	0.468	0.445	0.358
	10%	0.693	0.501	0.506	0.486
	15%	0.702	0.522	0.533	0.569
	20%	0.721	0.516	0.561	0.631

Table 1: Comparison of the NRMSEs based on KNN, LLS, BPCA and NN imputation.

References

[1] Abdi, H. 1994. A neural network primer. *Journal of Biological Systems*, 2:247–283.

[2] Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Sherlock, G., Chan, W.C., Greiner, T. C., Weisenburger, D.D., Armitage, J.O., Wamke, R. and Staudt, L.M., et al. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511.

[3] DeRisi, J.L., Iyer, V.R. and Brown, P.O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686.

[4] Kim, H., Golub, G.H. and Park, H. 2005. Missing value estimation for DNA microarray gene espression data: Local least squares imputation. *Bioinformatics*, 21:187–198.

[5] Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K. and Ishii, S., 2003. A Bsyesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19:2088–2096.

[6] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Bostein, D. and Futcher, B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast S. cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9:3273–3297.

[7] Troyanskaya, O., Cantor, M. Sherlock, G., Brown, P., Hastie, T., Tibshirant, R., Bostein, D. and Altman, R.B. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520–525.

Gene Expression Network Analysis

Serban Nacu[1]

1 Introduction

In [1] we introduced the GXNA algorithm in order to discover differentially expressed pathways in gene expression experiments. Biological knowledge was represented as a gene interaction graph, where nodes stand for genes and edges for interactions. GXNA searches the graph for small connected subgraphs with high scores for differential expression. When applied to several human cancer data sets, it finds pathways that were not identified by standard single gene analysis or by other pathway based methods.

Here we present several improvements that make the algorithm more powerful and easier to use.

2 Algorithms

The original search method was greedy expansion starting from several root nodes. We improve upon this by adding a Metropolis step that searches a larger section of the state space, while constraining the target subgraph to remain connected. We also speed up the greedy step by keeping a list of neighbors for each node and pre-sorting them each time the node scores are computed. Together, these improvements lead to more accurate reported pathways and reduced runtimes. A typical analysis, including the computation of familywise error rates, takes less than a minute on a normal desktop.

3 Software

We also present updates to the GXNA software. The program now has simplified inputs, and produces graphical output that can be visualized in any web browser. We implemented scores based on F-statistics, allowing the comparison of multiple phenotypes or conditions. We also added annotation files for several microarray platforms, and an updated gene interaction graph that tracks the type and source of each interaction.

The software is available for free download at `http://stat.stanford.edu/ serban/gxna`.

References

[1] Nacu, S., Critchley-Thorne, R., Lee, P. and Holmes, S. 2007. Gene Expression Network Analysis and Applications to Immunology. *Bioinformatics*, 23:850–858.

[1]Department of Bioinformatics, Genentech, South San Francisco, CA, USA. Email: `serbann@gene.com`

Discovery of DNA Copy Number Variation Using Shotgun Sequencing Data

Chao Xie,[1] Martti T. Tammi[2]

1 Introduction

DNA copy number variation (CNV) is an important type of genome variation. In human genome, 360 megabases (12% of the geome) were identified as CNV regions [3]. CNV are usually discovered using array Comparative Genomic Hybridization (aCGH) [2]. We developed a method called eCGH (or electronic CGH) to simulate aCGH in silico. eCGH uses whole genome shotgun sequencing data, either traditional capillary sequencing or the new 454 sequencing. eCGH can be readily applied to the large amount of existing sequencing data. Because shotgun sequencing is becoming cheaper and faster, eCGH will be a very useful tool for discovering CNV.

2 eCGH and CNV Discovery on Simulated Data

Our method, eCGH, works by simulating aCGH. It simulates whole genome DNA array using a known genome assembly. Shotgun sequencing data from two individuals are used to simulate the two sets of DNA samples in aCGH, because shotgun sequencing is a random sampling of DNA fragments. Sequencing data with sequencing coverage as low as 1/10x can be used. The hybridization process is simulated using sequence alignment. Log2 ratios of number of aligned shotgun reads from the two individuals were calculated for each sliding window on the genome assembly. CNV regions were discovered based on the log2 ratios. We tested our eCGH method on simulated data. More than 4 million simulations were performed. Parameters were optimized based the simulations. One example of the simulations is shown in Fig 1.

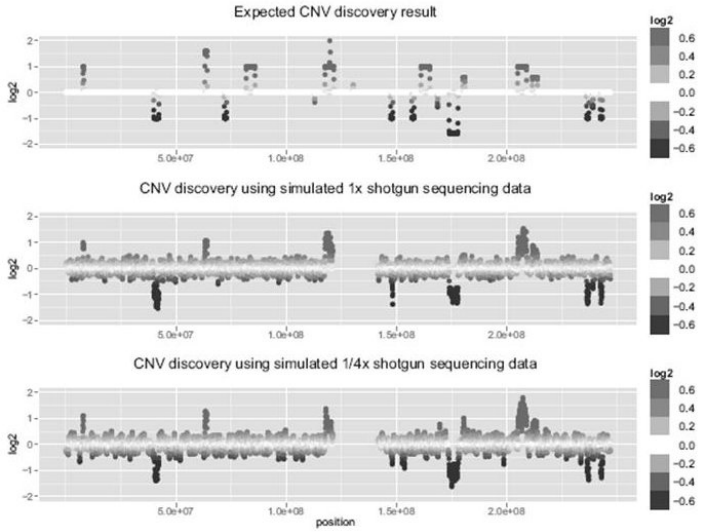


Figure 1: One example of CNV discovery on simulated shotgun sequencing data with different sequencing coverages.

ROC curves of the simulations are shown in Fig. 2. Higher sequencing coverage gives better performance. ROC for capillary sequencing data with 1x coverage has area under curve (AUC) 83.5%, while 1/4x coverage has AUC 78.5%. For the same sequencing coverage, 454 data performs better than capillary data. This also means that the emerging 454 sequencing method will provide much more input data for eCGH.

[1]Department of Biological Sciences, National University of Singapore, Singapore. Email: xie@nus.edu.sg
[2]Department of Biological Sciences, National University of Singapore, Singapore. Karolinska Institutet, Department of Microbiology, Tumor and Cell Biology, Stockholm, Sweden. Email: martti@nus.edu.sg

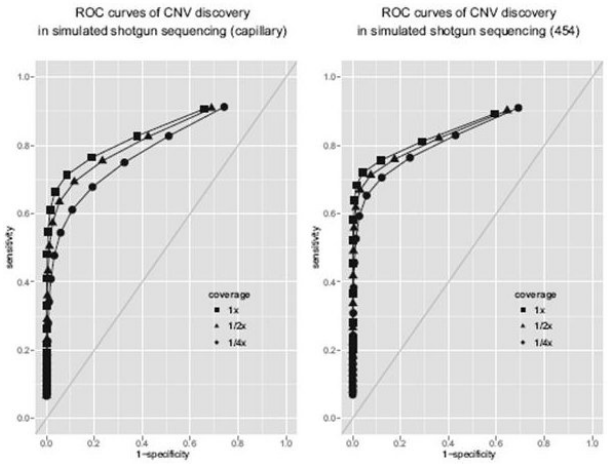


Figure 2: ROC curves of CNV discovery performance on traditional capillary and 454 shotgun sequencing data with differnet sequencing coverage.

3 CNV in Domestic Chickens

eCGH were used to discover CNV in three domestic chicken breeds (Broiler, Layer, and Silkie) [1]. The domestic chickens were sequenced at 1/4x coverage. The results are shown in Fig. 3. Most sliding windows on autosomes (chrom 1 to 28) have log2 around 0, or copy number ratio 1:1. The log2 ratios on chromosome Z are around 0 in Layer vs Silkie and Broiler vs Silkie, while around 1 in Broiler vs Layer and Broiler vs Silkie. This is because the sequenced Broiler is male, so with sex chromosomes ZZ, while Layer and Silkie are female, with WZ [1]. This result also shows that our eCGH works correctly.

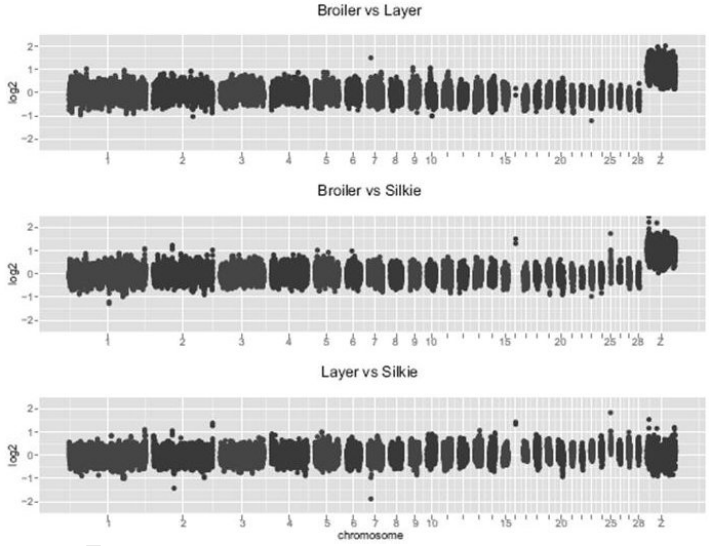


Figure 3: CNV among three domestic chicken breeds.

References

[1] International Chicken Polymorphism Map Consortium. 2004. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature*, 432:717–722.

[2] Pinkel, D. and Albertson DG. 2005. Array comparative genomic hybridization and its applications in cancer. *Nature Genetics*, 37 Suppl:S11–S17.

[3] Redon, R., Ishikawa, S., et.al. 2006. Global variation in copy number in the human genome. *Nature*, 444:444–454.

Genome-Wide High-Density ChIP-Chip Tiling Array Data Analysis in Fission Yeast

Juntao Li,[1] Lei Zhu,[2] Majid Eshaghi,[3] Jianhua Liu,[3]
R. Krishna Murthy Karuturi[1,4]

1 Introduction

ChIP-on-chip (also known as ChIP-chip) [1] is a technique that combines chromatin immunoprecipitation (ChIP) with microarray technology (chip). It allows the identification of binding sites of DNA-binding proteins in a very efficient and scalable way [2]. We propose a new procedure to analyze hi-density ChIP-chip tiling array data to characterize protein-DNA interaction. First, we identify the enriched signal regions or protein binding occupancies using moving window binomial analysis, classify the occupancies into three categories and process them separately. Single peak footprints were processed to get the peak position signifying binding location. Multi peak footprints are split into individual peaks and processed for binding locations. The flat binding occupancies were processed to summarize their overall strength. We applied our procedure to analyze the custom designed NimbleGen genome-tiling array data of ∼380K probes of fission yeast.

2 Moving Window Binomial Analysis

We applied the binomial test on sliding windows to identify regions with enriched signals using datasets resulted from ChIP-chip analysis. That is, we counted the number of probes that passed a threshold of median + 2.5 MAD (median of absolute deviation) in a sliding window of 9 consecutive probes in size (i.e., ∼300 bps) and obtained p-values resulted from binomial test. As a result, each probe was marked either "+" if it passed the threshold of median + 2.5 MAD with a p-value of < 0.001 or "-" if it did not. We next defined regions with > 4 consecutive "+" signed probes (i.e., ∼140 bps) as enriched region. These enriched regions indicate the presence of occupancies, whose footprint was defined using the same approach with the threshold of median + 1 MAD and the p-value of < 0.01. To this end, an occupancy could contain multiple enriched regions (M+2.5MAD, p-value < 0.001) within a single footprint (M+1MAD, p-value < 0.01).

3 Region splitting for Multi-peak

To avoid an occupancy covering multiple genomic features, the footprint between the two features was split at the trough of the occupancy profile. For doing this, we first assigned each probe "+" and "-", if R_i (the ratio R of probe i) was $> R_{i+1} + d$ and $< R_{i+1} - d$, respectively, where $d = \text{MAD}(\{\delta_i\})$, $\delta_i = |R_i - R_{i+1}|$ for $i = 1, 2, ..., n$, where n is the total number of probes between two features. The probe was assigned to "0" if $|R_{i+1} - R_i| < d$. After removing all probes with "0", the footprint was split between the opposite signs such as "–" and "+". After splitting, occupancies with footprint of less than 4 consecutive probes are removed. Peak position of occupancies was determined based on the smoothed profiles. That is, the position of probes with the maximal ratio within a footprint of occupancies is defined as the position of the peak for occupancies.

[1]Computational & Mathematical Biology, Genome Institute of Singapore, Singapore.
[2]MOE Key Laboratory for Biodiversity Science and Ecological Engineering and College of Life Sciences, Beijing Normal University, Beijing, P. R. China. Zhu lei was on attachment at the Genome Institute of Singapore during this work.
[3]Systems Biology, Genome Institute of Singapore, Singapore.
[4]Corresponding author. Email: karuturikm@gis.a-star.edu.sg

4 Peakedness Test for Binding Occupancies

Occupancies whose footprint covered more than 70% coding regions were tested for peakedness using kurtosis [3] within the window of footprint. The kurtosis value is based on the formula

$$K = \frac{\sum_{j=1}^{n}(j-u)^4 \times p_j}{(\sum_{j=1}^{n}(j-u)^2 \times p_j)^2}$$

where $u = \sum_{j=1}^{n} j \times p_j$ and $p_j = R_j / \sum_{i=1}^{n} R_i$, R_j is the probe ratio at jth position.

The occupancy was designated as having flat-shaped when its $K < 2.5$. Thus, intergenic or peaked occupancies are separated from flat occupancies.

Acknowledgments. We thank Edison T. Liu and Neil D. Clarke for their support during this work. We appreciate Jonghoon Lee, Chee Seng Chan, Atif Shahab for their valuable discussion. This work was supported by the Biomedical Research Council of Agency for Science, Technology and Research (A*STAR), Singapore.

References

[1] Buck MJ., Lieb JD. 2004. ChIP-chip: Considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments, *Genomics*, 83:349–360.

[2] Iyer VR., Horak CE., Scafe CS., Botstein D., Snyder M., Brown PO. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, 409(6819):533–538.

[3] Joanes, DN. Gill, CA. 1998. Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society (Series D): The Statistician*, 47(1):183–189.

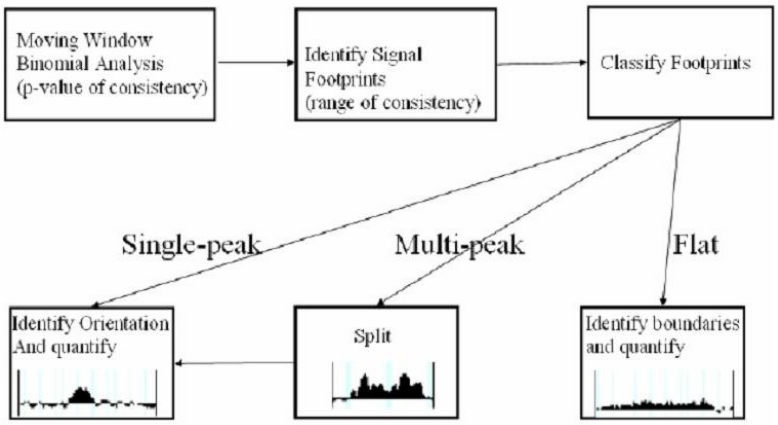


Figure 1: Workflow for the ChIP-chip data analysis.

The ANNOTATOR Software Environment: A Flexible Sequence Analysis Platform

Wong Chee-Hong,[1] Ooi Hong Sain,[2] Georg Schneider[3]

1 Introduction

The ever increasing amount of sequence data available in public databases is a mixed blessing. On the one hand it allows to build bridges between evolutionary distant sequences making it possible to predict function. On the other hand it makes the very task of finding these bridges time-consuming and tedious.

Clearly there is a need for automated and efficient methods to assist in the process of function discovery. Integrating these methods into a framework represents a huge advantage over developing ad-hoc solutions which are discarded once the specific question has been addressed.

2 The ANNOTATOR Software Environment

The ANNOTATOR software environment integrates a large number of external sequence analytic algorithms in a way that facilitates following a segment based approach to functional characterization.

Additionally a standardized interface allows to develop integrated algorithms with full access to internal data-structures and process-spawning capabilities. In this way it is possible to devise new methods for finding distant homologues by iterativelly collecting family members [1] or implement sophisticated clustering strategies [2, 3]. At the same time the use of a standardized framework allows to easily trace-back and validate decisions taken by the automated procedure. Additionally, results are available for further internal processing.

Here we present two implementations of integrated algorithms that demonstrate the above mentioned advantages.

3 Family-Searcher

The analysis of homology relationships within large superfamilies of protein sequences can be used to organize the sequence space of known proteins and elucidate the function and evolutionary origin of unknown ones [4].

The FAMILYSEARCHER algorithm, developed as an integrated algorithm within the ANNOTATOR software environment, uses an approach of fan-like iterative PSI-BLAST searches [5] which are coupled with sequence-analytic methods to detect compositional and repetitive pattern bias.

In this way it is possible to collect families with tens of thousands of members in an automated manner. The algorithm has been used in a number of sequence-analytic projects including the uncovering of the evolutionary relationship between classical mammalian lipases and human adipose triglyceride lipase [1].

4 Orthologue-Searcher

Orthology is seen as the relationship of descendance from a common ancestral sequence where sequence divergence followed speciation while paralogy describes a relationship where divergence followed gene duplication [6, 7, 8]. Correctly identifying orthologues greatly enhances the reliability of transferring functional information.

The algorithm implemented as an extension of the ANNOTATOR allows to identify orthologues of a given sequence using an adaptation of the Reciprocal-Best-BLAST-Hit approach. The rapdily rising number of sequenced organisms means that a typical Orthologue-Search will entail a large number of BLAST-searches including post-processing, a task that can only be fulfilled by an integrated system.

[1]Bioinformatics Institute, A*STAR, Biopolis, Singapore. Email: wongch@bii.a-star.edu.sg
[2]Bioinformatics Institute, A*STAR, Biopolis, Singapore. Email: ooihs@bii.a-star.edu.sg
[3]Bioinformatics Institute, A*STAR, Biopolis, Singapore. Email: georgs@bii.a-star.edu.sg

References

[1] Schneider, G., Neuberger, G., Wildpaner, M., Tian, S., Berezovsky, I., and Eisenhaber, F. Application of a sensitive collection heuristic for very large protein families: Evolutionary relationship between adipose triglyceride lipase (atgl) and classic mammalian lipases. *BMC Bioinformatics*, 7:164, 2006.

[2] Enright, A. J., Dongen, S. V., and Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30:1575–1584, 2002.

[3] Neuberger, G., Schneider, G., and Eisenhaber, F. pkaps: Prediction of protein kinase a phosphorylation sites with the simplified kinase-substrate binding model. *Biology direct*, 2:1, 2007.

[4] Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. Predicting function: From genes to genomes and back. *Journal of Molecular Biology*, 283:707–725, 1998.

[5] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[6] Koonin, E. V. An apology for orthologs—or brave new memes. *Genome Biology*, 2, COMMENT1005, 2001.

[7] Fitch, W. M. Homology a personal view on some of the problems. *Trends in Genetics*, 16:227–231, 2000.

[8] Jensen, R. A. Orthologs and paralogs—we need to get it right. *Genome Biology*, 2, INTERACTIONS1002, 2001.

A Bioinformatic and Transgenic Approach for Elucidating Tissue Specific Regulatory Elements

Sumantra Chatterjee,[1,2] Guillaume Bourque,[2] Thomas Lufkin[1,2,3]

Identifying the sequences that direct the spatial and temporal expression of genes remains a significant challenge in the annotation of vertebrate genomes. Precisely locating these sequences, which in many cases lie at a great distance from the transcription start site has been a major obstacle in deciphering the complete regulatory profile of a gene. The completion of a number of vertebrate genome sequences, as well as the concurrent development of genomic alignment, visualization, and analytical bioinformatics tools, has made large genomic comparisons not only possible but an increasingly popular approach for the discovery of putative cis-regulatory elements. The unprecedented coverage and resolution of the sequence data available makes it possible to compare sequence conservation between diverse species spanning the evolutionary spectrum.

Here we present a study which uses a genome alignment between extreme vertebrate species (teleost fish to humans) which last shared a common ancestor about 450 million years ago, to come up with a list of conserved noncoding elements to test for regulatory activity in important genes which control the specification and development of the sclerotome lineage in vertebrates (Bapx1, Sox6, Pax9, Foxc1a). We use available genome alignments from UCSC genome browsers and a fish transgenic system to locate and validate proximal as well as distal cis regulatory elements for these genes.

With the data available at our disposal we hope to form a robust yet flexible bioinformatic tissue specific "training set" that will be used as a template for insilico prediction of regulatory elements and help to decipher some common central theme regarding the regulatory mechanisms for genes critical for the development of a particular tissue.

References

[1] Adam Woolfe et al. (2005) Highly Conserved Non-Coding Sequences are Associated with Vertebrate Development. *PLoS Biology*, 3:116–130.

[2] Ahituv et al. (2005) Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny. *Human Molecular Genetics*, 14:3057–3063.

[3] Wasserman WW, Sandelin A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nature Review Genetics*, 4:276–287.

[4] Pennachio LA et al. (2006) In vivo enhancer analysis of human conserved non-coding elements. *Nature*, 444:499–502.

[1]Department of Biological Science, National University of Singapore.
[2]Genome Institute of Singapore.
[3]Corresponding Author. Email: `lufkin@gis.a-star.edu.sg`

Cellular Automata and Simulation of Biological Processes

Rosaura Palma-Orozco,[1] Jorge Luis Rosas-Trigueros[2]

1 Introduction

We are building computational models of biological and chemical systems using cellular automata and it is necessary to show the potentiality of this computational tool.

A cellular automaton CA (plural: cellular automata, [5, 1]), is a discrete model studied in computability theory, mathematics, and theoretical biology. It consists of a regular grid of cells, each in one of a finite number of states. The grid can be in any finite number of dimensions. Time is also discrete, and the state of a cell at time t is a function of the states of a finite number of cells (called its neighborhood) at time $t - 1$. These neighbors are a selection of cells relative to the specified cell, and do not change (though the cell itself may be in its neighborhood, it is not usually considered a neighbor). Every cell has the same rule for updating, based on the values in this neighborhood. Each time the rules are applied to the whole grid a new generation is created.

2 Some Rules

There are many specific rules that are used to simulate some biological processes, for example Rule 30 and Diffusion Rule.

Rule 30 is a one-dimensional binary cellular automaton rule introduced by Stephen Wolfram [5] in 1983. This rule is of particular interest because it produces complex, seemly random patterns from simple, well-defined rules. For instance, a pattern resembling Rule 30 appears on the shell of the widespread cone snail species conus textile, see section 3.

A two-dimensional cellular automata, where every cell takes states 0 and 1 and updates its state depending on sum of states of its 8 closest neighbors as follows. Cell in state 0 takes state 1 if there are exactly two neighbors in state 1, otherwise the cell remains in state 0. Cell in state 1 remains in state 1 if there are exactly seven neighbors in state 1, otherwise the cell switches to state 0, CA governed by such cell-state transition rule exhibits reaction-diffusion like pattern dynamics, so we call this Diffusion Rule [3], see Figure 1.

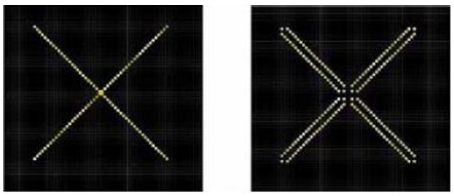


Figure 1: Diffusion Rule.

3 Examples and Applications

Cellular automata are now used to model several phenomena present in the our physical world [7, 4]. Some models can only be used to express a basic idea of a phenomenon, others are accurate enough to be used for prediction. Let us just list some examples of physical phenomena and CA that exhibit similar behavior:

[1]Department of Computer Sciences, Centro de Investigacio'n y de Estudios Avanzados, CINVESTAV - IPN, Mexico. Email: rpalma@math.cinvestav.mx

[2]Department of Graduate Studies, Escuela Superior de Co'mputo, ESCOM - IPN, Mexico. Email: jlrosas@ipn.mx

- Patterns on sea shells
- Growth of crystals especially patterns in snowflakes can be modeled by simple 2D
- Excitable media in biology (predator-prey dynamics)
- Fractal growth of biological organisms

Some living things use naturally occurring cellular automata in their functioning. Patterns of some seashells, like the ones in *Conus* and *Cymbiola* genus, are generated by natural CA (see Figure 2). The pigment cells reside in a narrow band along the shell's lip. Each cell secretes pigments according to the activating and inhibiting activity of its neighbor pigment cells, obeying a natural version of a mathematical rule.

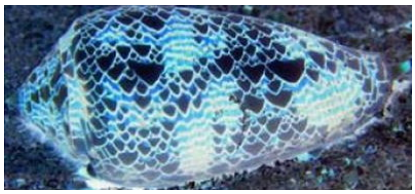


Figure 2: *Conus textile* exhibits a cellular automata pattern on its shell.

References

[1] Andrew Adamatzky. *Identification of Cellular Automata*. Taylor and Francis, London, 1994.

[2] Howard Gutowitz, editor. Cellular Automata: Theory and Experiment, 1991. Published as *Physica*, D45 (1990) Nos. 1–3, and as MIT press book.

[3] `http://parts2.mit.edu/wiki/index.php/IPN_UNAM_2006`.

[4] `http://parts.mit.edu/igem07/index.php/Mexico`.

[5] Stephen Wolfram. *Cellular Automata and Complexity: Collected Papers*. Addison-Wesley, 1994.

[6] Stephen Wolfram, editor. *Theory and Applications of Cellular Automata*. World Scientific, Singapore, 1986.

[7] Toffoli T., Margolus N. *Cellular Automata Machines: A New Environment for Modeling*, The MIT Press, 1987.

Genomic Analysis of Transcriptional Regulation by the Factor REST in Embryonic Stem Cells

Rory Johnson,[1] Galih Kunarso,[1] Christina Teh,[1] Kee-Yew Wong,[1] Kandhadayar G. Srinivasan,[1] Sarah S.-L. Chan,[1] R. Krishna Murthy Karuturi,[1] Leonard Lipovich,[1] Noel J. Buckley,[1] Lawrence W. Stanton[2]

1 Introduction

Pluripotency of Embryonic Stem (ES) cells is controlled by specific, interconnected gene regulatory networks. Understanding and reconstructing these networks will be important in future therapeutic applications of stem cells. One important player in this network is REST (RE1-silencing transcription factor), a transcriptional repressor involved in multiple developmental and disease processes [1]. The aim of this project is to understand the role of REST in pluripotency by comprehensively reconstructing its regulatory target genes in ES cells. We have used complimentary high-throughput genomic methods to (a) map the recruitment of REST to the genome of ES cells, and (b) measure the functional outcome of this recruitment on target genes.

2 Results

An unbiased mapping of REST binding regions in mouse ES cells was performed by the sequencing-based method, ChIP-PET (Chromatin Immunoprecipitation coupled to Paired-End diTagging). This approach detects REST binding events based on the identification of clusters of overlapping sequences from immunoprecipitated genomic DNA. Using an empirical cut-off strategy we identified 2460 high-confidence REST binding sites of 5 or more overlapping fragments (moPET5+) in ES cells.

REST has been an important model for the study of transcriptional regulation because of the length and specificity of its 21bp recognition element, which has facilitated computational identification of target genes [2]. Using the de novo motif finding algorithms MEME and WEEDER, we were able to reconstruct the canonical RE1 motif from the regions underlying ChIP-PET clusters. Nevertheless, the majority of bound loci do not contain a high-quality RE1—in fact only 678 contain a high-quality RE1. Instead, most sites contain either degenerate, full-length RE1 motifs, or various combinations of RE1 half-sites (Table 1). Therefore, REST recruitment is less sequence-dependent than previously supposed, suggesting that RE1-independent sequences, chromatin state, or other transcription factors contribute to targeting REST in ES cells.

Motif	# E14 moPET5+	%
Full RE1	1698	69.02
Left & Right	200	8.13
Left-only	115	4.67
Right-only	246	10.00
No RE1	201	8.17
Sum	2460	

Table 1: Sequence analysis of REST PET regions. The numbers of PET clusters having various RE1 sequence configurations are shown. Motifs were identified using a position-specific scoring matrix (PSSM) for the RE1 with a weak cutoff threshold, as well as individual RE1 half sites. Left & Right refers to non-canonical combinations of half-sites: altered spacer distance, convergent, divergent and inverted.

We also investigated the functional output of REST recruitment on target gene expression. Global gene mRNA levels from control cells, as well as those where REST function was ablated by means of

[1]Centre for the Molecular Basis of Behaviour, Institute of Psychiatry, King's College, London SE5 8AF, UK.
[2]Genome Institute of Singapore, 60 Biopolis Street, #01-02, Singapore 138672. Email: {stantonlw, johnsonrb, galihk@gis.a-star.edu.sg

a dominant-negative construct (DN:REST), were measured using Illumina BeadChip microarrays. We identified 441 genes in ES which change significantly under these conditions. Comparison of regulated genes to REST PET locations revealed that repressed genes are highly enriched for REST PET clusters within 100kb, while activated genes are not (Figure 1). Repressed genes predominantly recruit REST to their proximal promoter region (Figure 2).

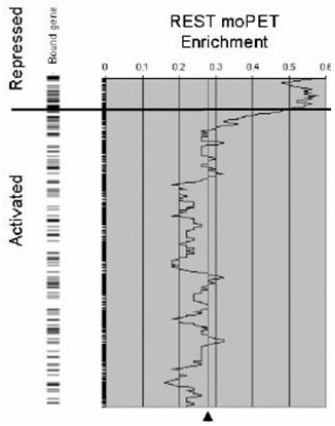


Figure 1: REST target genes are enriched for REST PET clusters. The vertical axis represents ranked genes whose expression significantly changes with DN:REST (the bar divides repressed from activated genes). The horizontal axis represents the mean of the enrichment of REST PET within 100kb of the genes, for a sliding window. The arrowhead indicates the genomic background enrichment.

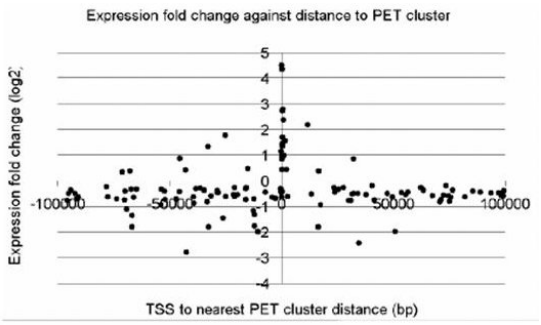


Figure 2: Repressed target genes are predominantly regulated by promoter-binding of REST. Y-axis represents log fold change in gene expression upon DN:REST—genes which are repressed by REST have positive values; x-axis represents distance of nearest REST moPET5+ cluster to gene transcriptional start site.

3 Conclusions

We have performed a comprehensive, whole-genome analysis of REST, an important component of the ES pluripotency gene network. Our findings have yielded many promising new target genes of REST which may hold the key to understanding its function. The data suggest that sequence considerations alone are probably insufficient to reconstruct genomic binding in a given cell type. Future studies will apply similar techniques to understanding the role of chromatin state and other transcription factors in defining recruitment and regulation.

References

[1] Ooi, L. and Wood, I.C. 1998. Chromatin crosstalk in development and disease: Lessons from REST. *Nature Reviews Genetics*, 8(7):544–554.

[2] Johnson, R. et al 2006. Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res.*, 34(14):3862–3877.

Genetic Characterization and Population Structure of Arabian Tahr (*Hemitragus jayakari*) based on Microsatellites Analysis

Mohammed A. Khidhir, K. Praveen Kumar, and Marwa Al-Aseer[1]

1 Introduction

We report here the first study on the genetic structre of the Arabian Tahr (*Hemitragus jayakari*) population, which is unique to the mountain of the South East Asia (UAE & Oman). This species is listed as vulnerable and endanger by the World Conservation Union (IUCN). Since no microsatellite markers were isolated from this species so far, we selected 35 microsatellite markers from cattle (Bishop et al., 1994), sheep and goats (A.Kotze et al., 2004; A.M.Martinez et al., 2004). Based on the analysis of 11 polymorphic microsatellite markers, the study revealed that Arabian Tahr need appropriate genetic management for their conservation.

2 Method

DNA was extracted by using a QIAamp DNA midi kit (QIAGEN). All PCR amplifications were conducted in ABI 9700 Thermocycler. The amplicon is Genotyped by using Genotyper 2.0 Version (ABI 3100 Genetic Analyzer; Applied Biosystems).

3 Results

A total of 28 microsatellite markers were isolated, among them 11 were polymorphic; one polymorphic marker is shown in Fig. 1. Standard diversity indices of Arabain Tahr were calculated by using ARLIQUENE. The Hardy-Weimberg (HW) equilibrium was calculated by using GENEPOP as shown in Table 1. Relatedness between Individuals shown in Table 2 was calculated by using the IDENTIX software.

We found Average gene diversity over loci : 0.484885 +/- 0.261360.

4 Discussions

Genetic analyses for observed heterozygosity (Ho), expected heterozygosity (He), mean number of alleles and number of alleles per locus were calculated by ARLEQUINE software (Laurent Excoffier et al., 2006). Hardy-Weimberg (HW) equilibrium is calculated by using the GENEPOP software v. 3.1c (Raymond & Rousset, 1995), which applies the chain method of Monte Carlo Markov (Guo & Thompson, 1992). Relatedness between individuals is calculated by using IDENTIX software (Khalid et al., 2002). Among 28 sets of isolated microsatellite markers, 11 markers were polymorphic and average gene diversity over loci in the population is 0.484885 +/- 0.261360 which is moderate. Four markers (McM527, BM1258, INRABERN172 and ETH225) showed significant deviation from the Hardy-Weimberg (HW) equilibrium. This indicates that proper management has to be taken for the conservation of the Arabian Tahr.

Acknowledgments. We thank H.H. Shaikh Khalifa Bin Zayed Al Nahyan, President of the UAE, for his endless support for this project.

[1]. Management of Nature Conservation, Department of the President's Affairs, AL-AIN, U.A.E.

References

[1] A.Kotze, H. Swart, J.P. Grobler (2004). Genetic profiles of the Kalahari Red goat breed from South Africa. *J. South African Journal of Animal Science* 2004, 34.

[2] Bishop MD, Kappes SM, Stone RT (1994). Genetic linkage map for cattle. *J. Genetics*, 136, 619–639.

[3] Khalid Belkhir and Vincent Castric (2002). IDENTIX: Software to test for relatedness in a population using permutation methods. *J. Molecular Ecology*, Notes 2, 611–614.

[4] Laurent Excoffier, Guillaume Laval (2006). ARLEQUINE: An integrated software package for population genetics data analysis.

[5] M. Raymond, F. Rousset (1995). Genepop: Population genetics software for exact tests and ecumenicism. *J. Heredity*, 86, 248–249.

Locus	Num of alleles	Obs. Het.	Expt. Het.	P-val.
HEL10	3	0.7	0.63495	0.8316
McM527	4	0.28	0.40869	0.0002
MAF65	4	0.46939	0.60446	0.0799
ILSTS011	3	0.48	0.52444	0.4259
BM1258	2	0.2	0.36848	0.0023
INRABERN172	2	0.28	0.42424	0.0209
TGLA53	3	0.22	0.23576	0.1331
ILSTS87	2	0.36	0.49778	0.0817
ETH225	2	0.32	0.47596	0.0328
OarAE54	3	0.4	0.50343	0.2075
SRCRSP5	4	0.68	0.6796	0.8895
Mean	**2.909**	**0.39904**	**0.48707**	
s.d	**0.831**	**0.16994**	**0.12725**	

Table.1; Standard diversity indices of Arabain Tahr.

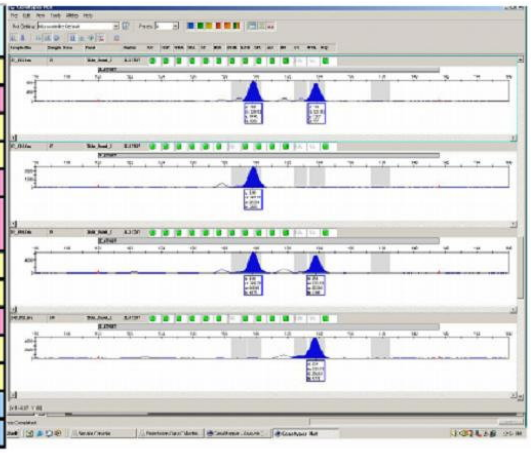


Fig. 1; Genotypes of the ILSTS87 marker.

	28F	29F	30F	31F	32F	33F	34F	35F	36F	37F	38F	39F	40F	41F	42F	43F	44F	45F	46F	47F	48F	49F
1M	18%	60%	39%	38%	69%	29%	26%	56%	51%	65%	48%	69%	43%	69%	59%	51%	92%	45%	65%	29%	72%	30%
2M	72%	54%	91%	78%	52%	76%	58%	69%	62%	82%	85%	77%	68%	60%	59%	83%	54%	72%	73%	70%	30%	61%
3M	57%	54%	80%	74%	45%	67%	60%	71%	67%	72%	93%	52%	94%	50%	60%	50%	57%	62%	71%	70%	35%	64%
4M	74%	56%	82%	74%	47%	77%	67%	74%	58%	76%	81%	72%	55%	55%	45%	78%	50%	68%	64%	78%	33%	60%
5M	74%	71%	74%	79%	63%	64%	43%	67%	66%	81%	80%	70%	61%	52%	66%	92%	62%	59%	83%	56%	45%	55%
6M	78%	72%	79%	84%	56%	63%	61%	69%	77%	64%	83%	57%	71%	57%	52%	72%	68%	57%	80%	69%	48%	59%
7M	50%	37%	75%	61%	33%	70%	73%	75%	59%	67%	57%	59%	94%	49%	67%	52%	50%	72%	46%	74%	44%	62%
8M	77%	66%	71%	88%	62%	59%	46%	72%	82%	66%	70%	66%	53%	49%	41%	89%	57%	56%	71%	64%	49%	56%
9M	47%	75%	58%	76%	65%	38%	41%	76%	59%	63%	71%	49%	61%	52%	63%	63%	75%	37%	73%	58%	68%	54%
10M	58%	52%	87%	77%	51%	83%	81%	82%	74%	73%	73%	55%	83%	69%	66%	65%	62%	72%	62%	90%	50%	66%
11M	62%	37%	69%	70%	27%	79%	78%	64%	45%	47%	55%	34%	59%	48%	39%	49%	32%	62%	34%	86%	27%	76%
12M	71%	49%	72%	68%	27%	65%	60%	56%	49%	39%	61%	47%	60%	41%	28%	56%	46%	62%	52%	67%	28%	61%
13M	67%	69%	84%	79%	62%	62%	93%	77%	68%	70%	81%	71%	92%	73%	68%	68%	79%	67%	90%	66%	62%	61%
14M	77%	50%	97%	84%	42%	88%	75%	77%	70%	68%	82%	62%	85%	62%	53%	68%	52%	77%	63%	87%	34%	68%
15M	49%	64%	53%	52%	74%	51%	32%	56%	63%	77%	57%	85%	39%	68%	49%	78%	75%	65%	68%	37%	65%	27%
16M	73%	59%	74%	66%	38%	79%	74%	91%	71%	59%	78%	59%	70%	54%	41%	54%	56%	59%	58%	86%	54%	54%
17M	55%	47%	78%	68%	60%	85%	69%	78%	69%	82%	65%	75%	71%	76%	65%	72%	61%	81%	56%	78%	55%	57%
18M	59%	50%	69%	52%	39%	61%	65%	70%	66%	63%	65%	60%	62%	47%	37%	56%	57%	61%	63%	66%	47%	40%
19M	44%	36%	82%	68%	54%	71%	62%	52%	44%	61%	63%	49%	69%	68%	60%	52%	45%	72%	49%	71%	28%	75%
20M	26%	42%	50%	41%	54%	50%	48%	52%	43%	72%	48%	74%	39%	67%	58%	62%	65%	56%	54%	45%	50%	43%
21M	56%	66%	71%	61%	71%	74%	63%	68%	79%	95%	71%	76%	59%	72%	72%	79%	69%	79%	72%	67%	59%	46%
22M	64%	47%	63%	46%	30%	68%	77%	74%	73%	54%	67%	59%	53%	51%	29%	52%	51%	48%	61%	74%	42%	40%
23M	42%	65%	59%	63%	51%	47%	48%	53%	58%	66%	64%	50%	51%	49%	65%	64%	81%	43%	70%	48%	55%	42%
24M	66%	61%	87%	81%	68%	76%	53%	67%	70%	86%	80%	75%	72%	75%	72%	85%	65%	77%	76%	66%	46%	60%
25M	51%	84%	62%	62%	70%	55%	59%	68%	76%	78%	76%	57%	58%	68%	66%	59%	79%	63%	75%	67%	71%	48%
26M	37%	65%	68%	63%	56%	45%	50%	52%	63%	67%	69%	41%	71%	57%	73%	49%	74%	54%	67%	55%	51%	49%
27M	51%	59%	71%	71%	67%	44%	44%	70%	46%	57%	75%	69%	49%	62%	37%	65%	65%	47%	69%	61%	52%	58%

Table. 2; Relatedness between Male (M) & Female (F) in the population of Arabian Tahr

Guided-Discovery of Motifs for Peptide Binding Prediction

Menaka Rajapakse,[1,2] Lin Feng[3]

1 Introduction

Identifying T cell epitopes plays an important role in designing epitope-based vaccines. Computational prediction of peptides that bind to MHC class II molecules can effectively reduces the wet-lab experiments for identifying T cell epitopes. Peptides that bind to class II MHC molecules have broad length variations. Therefore, discovery of binding motifs in unaligned peptide sequences remains a fundamental problem in class II MHC peptide binding prediction. Often, prior knowledge such as anchor position specific residues or information acquired from experimental motifs is used to obtain an alignment prior to devising a predictive methodology. Probabilistic methods use only positive samples to discover motifs in unaligned peptide sequences. In this paper, we present a new approach which uses both positive and negative samples as well as motifs predicted by computational methods or by experimental methods to discover motifs for predicting binding peptides. Motif discovered by the proposed approach demonstrated better predictive performance for the datasets tested, compared to the performance obtained with the motifs derived from probabilistic and experimental methods.

2 Materials and Methods

We develop a new motif discovery approach guided-discovery for identifying a motif with better predictive power in unaligned class II MHC peptide sequences. The guided-discovery approach is summarized by the following steps: (1) Obtain computationally or experimentally predicted guiding motifs, (2) Construct profile matrices and scale for uniformity in representation, (3) Define objective functions representing the solution space, (4) Apply an optimization algorithm to determine the putative motif population, (5) Select the best solution from a putative motif solution population.

The computationally predicted guiding motifs were derived by using MEME [1] and Gibbs sampler approaches [3], whereas experimentally motifs were obtained from literature [5]. Motifs described as regular expressions were then represented as profile matrices. The objective functions, O_1 and O_2 were defined as below to capture the position specific information given in the computationally/experimentally predicted motifs and to improve the discrimination between binders (positives) and non-binders (negatives).

$$O_1 = \sum_g |\hat{Q} - Q(m(g))|;\ O_2 = FN + \kappa FP$$

where $\hat{Q}$ denotes the estimated position specific scoring matrix (PSSM) of the motif, and $Q(m(g))$ is the PSSM representation of the guiding motifs. The summation of O_1 is taken over all the guiding motifs. FN and FP represent false negatives and false positives and the factor κ is used to scale the ratio between binders and non-binders. An optimization algorithm [2] which is able to optimize (minimize) the multiple objectives, O_1 and O_2 simultaneously was used to obtain a solution population. The final phase involved choosing the best motif solution from the resulting solution population.

We applied the guided-discovery approach to DRB1*0401 dataset obtained from IEDB database as described in [4] and I-A^{g7} dataset described in [5]. The DRB1*0401 dataset of comprises of 209 binders and 248 non-binders. In order to simulate an independent evaluation, three mutually exclusive training sets were formed from the DRB1*0401 dataset. Two of which carrying only 50 binding peptides of 9aa or longer in length were assembled to train MEME and Gibbs sampler and retrieve the respective predicted motifs. The third set was used to train guided-discovery approach together with 124 non-binders. The

[1]Institute for Infocomm Research, Singapore. Email: menaka@i2r.a-star.edu.sg
[2]School of Computer Engineering. Nanyang Technological University, Singapore.
[3]School of Computer Engineering, Nanyang Technological University, Singapore. Email: asflin@ntu.edu.sg

remaining peptides, 59 binders and 124 non-binders were used as the testing dataset. Three-fold cross-validation was adopted and the average of the results was reported as the final performance value. In the case of I-A^{g7}, the training dataset consists of 438 binders and 134 non-binders. As subsets of the I-A^{g7} training dataset have been used to identify experimental motifs, we used an independent dataset for testing. The testing dataset comprises of 112 binders. Due to the lack of non-binders and to perform an unbiased evaluation of performance, 25 data sets, each comprising 112 non-binders were randomly generated, and combined with the binder dataset. The performance was reported by averaging the results over the 25 datasets. The area under receiver operator characteristics (AUC) was used to measure the performance. During each experiment, the PSSM which scored the highest AUC for the training set from the guided-discovery approach is retained and subsequently used to estimate the performance of testing data.

3 Results and Discussion

We applied the proposed guided-discovery approach to two different datasets. For testing the effect of computationally predicted motifs in guiding the motif discovery process, we used motifs predicted by two different probabilistic methods. We also tested the performance of the proposed method when guided by experimentally determined motifs. The performances of the proposed guided-discovery approach on the aforementioned instances are illustrated in Figure 1. In both cases, the predictive power of guided-discovery approach exceeded the performance of the motifs discovered by probabilistic methods and by experimental methods. Significant improvement in the performance was observed for the I-A^{g7} dataset guided by the experimental motifs. However, the integration of low-performing motif information predicted by computational means resulted in a marginal improvement in the overall performance of the DRB1*0401 dataset. We conclude that guided-discovery is an efficient approach that is able to improve the performance of a prediction algorithm built on motifs detected by experimental methods or by computational methods.

References

[1] Bailey, T.L. and Elken, C. 1995. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21:51–80.

[2] Deb, K., Pratap, A., Agrawal, S. and Meyarivan, T. 2002. A Fast and Elitist Multiobjective Genetic Algorithm:NSGA-II. *IEEE Trans on Evolutionary Computation*, 6(2):82–197.

[3] Nielsen, M., Lundegaard, C., Worning, P., Hvid, C. S., Lamberth, K., Buus, S., Brunak, S. and Lund, O. 2004. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*, 20:1388–1397.

[4] Nielsen, M., Lundegaard, C. and Lund, O. 2007. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics*, 8:238.

[5] Rajapakse, M., Schmidt, B., Lin, F. and Vladimir, B. 2007. Predicting peptides binding to MHC class II molecules using multi-objective evolutionary algorithms. *BMC Bioinformatics*, 8:459.

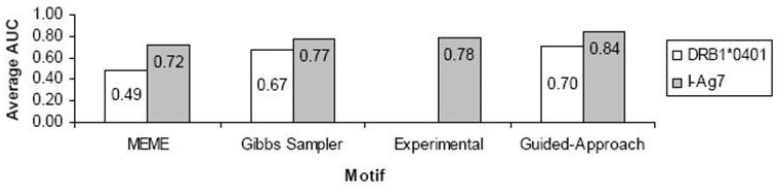


Figure 1: Illustration of the average AUC values obtained for the testing data sets of DRB1*0401 and I- A^{g7}. Computationally predicted motifs were used to guide the DRB1*0401 motif discovery process. The motif discovery of the I-A^{g7} was guided by the experimentally determined motifs. The average performance estimated for the I- A^{g7} testing dataset from number of experimentally determined motifs is given.

Mining for Domain Dependency Sets from Protein Interactions

Xiao-Li Li and See-Kiong Ng[1]

1 Introduction

Many protein interactions are known to be mediated by the underlying domain interactions among the proteins, and various computational methods have been developed to infer domain interactions. However, all the existing techniques adopt pairwise domain interaction models that only consider domain pairs and/or domain combination pairs. These approaches do not take into consideration the interactions occurring between multiple (more than two) domains, such as interactions in multi-protein complexes, or those that required third-party domains for activation. In this paper, we mine for such sets of multiple domains required for interactions. We denote them as Domain Dependency Sets (DDS). The DDS's are non-reducible sets of essential interacting domains with high support in protein interaction data; an exclusion of any domain member in a DDS will result in non-interacting protein sets. Our results show that statistically significant and biologically meaningful DDS's can be discovered from protein interaction data.

2 Problem Formulations

Mining domain-domain interactions was the earliest interaction mining efforts for shedding light on the underlying mechanisms of protein interactions and revealing potential protein interactions. Sprinzak et al. [1] were amongst the first to attempt to characterize protein interactions using domains in InterPro [2]. These early works had focused on pairwise interactions between individual domains.

Such pairwise models for domain-domain interactions do not represent the entire interaction rules, for there are many protein interactions known to mediate by multiple (more than two) domains. For example, protein phosphorylation, an essential process for many interactions and enzymatic reactions, is carried out by kinases which may not be involved in the direct interactions between two domains. In the signaling process by the receptor tyrosine kinase (RTK), two tyrosine kinase domains are needed to phosphorylate each other before the SH2 domains bind to them. Another similar scenario is the phosphorylation of protein inhibitors by a kinase to bind phosphoprotein phosphatase. Often, the kinases and enzymes that regulate interactions are found as individual proteins but many have been found fused to their target proteins (a phenomenon called gene fusion [3]). Such evolved proximity to their targets is speculated to increase the efficiency of these enzymes.

In addition to the enzymatic cases where the enzymatic domains are required as activators in protein-protein interactions, additional domains could also be needed to assist proteins to adopt the right structural conformation for interactions. Such domains function like molecular chaperones that assist in the folding of proteins and their assembly into complexes though they are not involved directly in the interactions. Multiple domains could also bind in a cooperative manner such that the absence of any domain would lead to non-interactions. Additional domains may also be required to serve as adaptor proteins that facilitate interactions between two proteins. Again, these multiple interacting domains may be found to exist individually in complexes or as fused products in interaction between two proteins. The requirement for multiple domains provides an added level of control for cell to regulate its protein-protein interactions.

Han et al. [4, 5] recently designed a probabilistic framework that takes pairs of domain combination sets instead of domain pairs as the basic units for protein interactions. In their work, members of each domain combinations are fixed. This is not the case in nature as domain combinations can be shuffled due to gene fusion (which glued multiple domains together) or gene fission (which split a multi-domain protein into individual components) genetic events. Interactions will occur as long as all essential domains are present regardless of the distributions of domains in different proteins. In this paper, we therefore consider a *dependency* set model for multi-domain interactions that do not assume pairwise relationships

[1]Institute for Infocomm Research, Singapore. Email: { xlli, skng}@i2r.a-star.edu.sg

between interacting protein domains or fixed domain combination sets. More specifically, this paper studies the following problem:

> **Protein Interaction mining of Domain Dependency Sets (DDS):** Find high support non-reducible sets of multiple domains (> 2 domains) that are deterministic of protein interactions.

We give an example to illustrate a DDS. One DDS discovered in our experiments is the domain set {PF00076, PF02847, PF04851}. It has high support as it was found to occur many times (8 times) in interacting protein pairs and complexes. It also occurred very sparsely in our non-interacting protein data sets, while all its subsets {PF00076, PF02847}, {PF00076, PF04851} and {PF02847, PF04851} do not interact-they occurred significantly more frequently in the non-interacting protein dataset than the interacting protein dataset.

The non-reducibility requirement means that any subset of a DDS cannot be an interacting set. This ensures that the protein interactions mediated by a DDS are not caused by any of subset of its member domains but by all its domains " concert". In this way, the DDS's provide the biologists with important information, such as what are all the necessary actors (domains) required to realize a protein interaction.

For the first time, we can capture the full range of dependency relationships of multiple interacting domains that underlies a protein-protein interaction; our DDS model does not assume fixed distributions among the various protein players in all the protein interactions that a DDS mediates. In other words, the key difference between a DDS and a conventional domain combination pair [4, 5] lies in that the domains in a DDS can be from different distributions or structures in different interactions that the DDS mediates. For example, given a DDS {A, B, C}, the triplet can be found pairing up in different combinations in pairs of interacting proteins ({-A-, -B-C-}, {-B-, -A-C-}, {-C-, -A-B-}) and/or alone in separate proteins in protein complexes ({-A-, -B-, -C-}). The only restriction of a DDS is that not all the domains can be from the same protein.

3 Results and Conclusions

For this new data mining problem, we have proposed a novel interaction mining algorithm with pruning strategies to ensure efficiency. We found that DDS model provide a better domain interaction model for protein interaction prediction than the conventional pairwise model. For example, the average interacting probabilities of all the pairwise subsets of 3-domain dependency sets extracted is $\sim$26%. However, the average interacting probabilities increases to 81% when we use 3 domains together. This show that even though the individual pairwise domain-domain subsets were useless for predicting interactions ($\sim$26% accuracy), a combination of these domains for predictions could reveal potential new interactions much more accurately. Mining domain dependency sets is a new and important approach for dissecting the underlying general mechanisms for protein-protein interactions. Instead of pairwise interacting domains and domain combinations, dependency sets containing multiple domains (> 2) deterministic of protein interactions were mined from protein interaction data including complex data. The resulting DDS's mined from yeast protein interaction data were found to be statistically significant and contained biologically meaningful information. We found that some of the extracted DDS's correspond to known biological knowledge, indicating that many of the other DDS's could correspond to potential novel discoveries worth further investigations by biological experiments.

References

[1] E. Sprinzak and H. Margalit. Correlated Sequence-signatures as markers of protein-protein interaction. *Journal of Mol. Biol.*, 311:681–692, 2001.

[2] R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, 29:37–40, 2001.

[3] A. J. Enright, I. Illiopoulos, N. C. Kyrpides, and C.A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:86–90, 1999.

[4] D. Han, H. Kim, J. Seo, and W. Jang. Domain combination based probabilistic framework for protein-protein interaction predication. *Genome Informatics*, 14:250–259, 2003.

[5] D. Han, H. Kim, W. Jang, and S. Lee. Domain combination based protein-protein interaction possibility ranking method. *BIBE2004*, pages 434–441, 2004.

SHRiMP: The Short Read Mapping Package

Stephen Rumble,[1] Michael Brudno[1,2]

1 Overview

Next generation sequencing technologies are revolutionizing the study of variation among individuals in a population. The ability of sequencing platforms such as AB SOLiD and Illumina (Solexa) to sequence one gigabase or more in a few days has allowed us to re-sequence a human genome (at 1x) in about 10 days on a single machine.

Here we present the Short Read Mapping Package (SHRiMP): a method for mapping very short reads to a genome. Our method includes 1) a spaced k-mer filtering technique, 2) a very fast, vectorized implementation of the Smith-Waterman algorithm, 3) separate full color-space and letter-space alignment approaches, and 4) computation of false discovery statistics for hits. We show the results of using SHRiMP for mapping reads to the *Ciona savingyi* reference genome, confirming the high heterozygosity of this genome for a second individual, and quantifying the sequencing error rate in the AB SOLiD datasets available to us. SHRiMP is freely available at `http://compbio.cs.toronto.edu/shrimp`.

2 Mapping Strategy

The algorithm starts with a rapid k-mer hashing step to localize potential areas of similarity between the reads and the genome. All of the spaced k-mers [1] present in the reads are indexed and for each k-mer in the genome, all of the matches of that particular k-mer among the reads are found. The approach of indexing the reads, rather than the genome has several advantages: First, it allows us to control memory usage, as our algorithm always needs memory proportional to the size of the genome, while the large set of short reads can be easily divided between many machines in a compute cluster. Secondly, our algorithm is able to rapidly isolate which reads have several k-mer matches within a small window by using a circular buffer to store all of the recent positions in the genome that matched the read. While this approach lowers the likelihood that k-mer matches are collinear, compared with the approach of Rasmussen et al [2], the lower per k-mer cost of computation combined with a fast (vectorized) implementation of the Smith-Waterman algorithm makes this approach advantageous for shorter k-mers.

If a particular read reaches a threshold number of k-mer matches within a given window of the genome, we execute a vectorized Smith-Waterman step to score and validate the similarity. The top n highest-scoring regions are retained, filtered through a full backtracking Smith-Waterman algorithm, and output at the end of the program if their final scores meet a specified threshold. The running time of the SHRiMP algorithm at various parameters is summarized in Table 1.

3 Color-Space Alignment

The AB SOLiD sequencing technology introduced a novel dibase sequencing technique, which reads overlapping pairs of letters and generates one of four colors (typically labeled 0-3) at every stage. The exact combinations of letters and the colors they generate is shown in Figure 1A. The sequencing code can be thought of as a finite state automaton (FSA), in which each previous letter is a state and each color code is a transition to the next letter state. This automaton is demonstrated in Figure 1B. We implement an algorithm for aligning color space reads in letter space. Our key observation is that while a color-space error causes the rest of the sequence to be mistranslated, the genome will match one of the other three possible translations. Consequently, we adapt the classical dynamic programming algorithm to simultaneously align the genome to all four possible translations of the read, allowing the algorithm to move from one translation to another by paying a "crossover", or sequencing error penalty. If one wishes for a probabilistic interpretation of the algorithm, one can consider the FSA in Figure 1B to be a Hidden Markov Model, where the letter is the hidden state, and the color-space sequence is the output of the model. By taking the cross product of this HMM with the standard pair-HMM associated with the Smith-Waterman algorithm, we can allow all of the typical alignment parameters, including the error

[1]Department of Computer Science and [2] Banting and Best Department of Medical Research. University of Toronto. Email: {rumble, brudno}@cs.toronto.edu

penalty, to be probabilistically motivated as the logarithm of the probability of the event, and trained using the Expectation-Maximization algorithm. It is notable that our approach handles not only matches, mismatches, and sequencing errors, but also indels.

4 Computing P-Values

As part of the SHRiMP package we provide the user with several utilities to simplify SNP discovery, including the ability to compute p-values as the probability that a match equally good or better would occur in a genome of equal length where at every position each nucleotide can be randomly selected with probability 0.25. We base this on the number of k-mer strings that have fewer then the observed number changes. The number of such strings grows by a factor of four with every mismatch and, (for color-space) crossover. For indels, the number of strings grows by a factor proportional to 2 times the length of the string: the 2 corresponds to having either an insertion or a deletion, and the length represents the possible positions for the indel. Now let the total number of strings be Z; we compute the p-value as $1 - (1 - Z/4^k)^{\text{genome\_length}}$.

5 Application to *C. savingyi* Resequencing

We use SHRiMP to explore the ability of AB SOLiD's read sequencing technology to capture the polymorphisms present in the sea squirt (*Ciona savingyi*) organism. The amount of variation present is nearly 50-fold higher than in human, making it one of the most difficult organisms for variation detection. Our results (Table 2) illustrate that the mapping of short reads, even in the presence of insertion and deletions, is feasible in the case of the highly polymorphic genome. Furthermore, the nature of AB SOLiD's dibase sequencing methodology allows us to differentiate between SNPs present in the genome and sequencing errors (on the assumption that the reference genome is accurate). Our results confirm the previously observed high polymorphism rate within the *Ciona savingyi* genome, and estimate the per-position error rate of the AB SOLiD technology as 2–3%, concordant with ABs own estimate.

References

[1] Ma B, Tromp J, Li M. PatternHunter: Faster and more sensitive homology search. *Bioinformatics*, 18(3):440–445, 2002.

[2] Rasmussen K., Stoye J., and Myers E.W. Efficient q-gram filters for finding all e-matches over a given length. In: *Proc. 9th Conf. on Computational Molecular Biology*, Boston, MA, 2005.

(A)

	A	C	G	T
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0

(B)

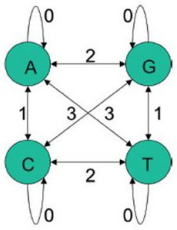


Figure 1: The dibase sequencing alphabet of AB SOLiD, presented as A, a transition matrix; and B, a Finite State Automaton. For example, and A followed by an A generates a "0" color, a C followed by a T a "2".

K-mer	7	8	9	10	12
% in SW	45%	25%	12%	7%	3%
Time (S)	2066	520	255	195	205

Table 1: The running time of SHRiMP for mapping 11,200 25-color reads to the *C. Savingyi* genome (180Mb) with various seed sizes (2 seeds are required to start an alignment). % in SW is the percent time spent within the vectored SW.

	p<.05	p<.01
% Reads mapped	20%	9%
SNP rate	.039	.024
Indel rate	.004	.003
Error rate	.024	.020

Table 2: Variation and error rates based on mapping 22 million *Ciona savingyi* reads from a second individual to the reference genome.

COCAW: Comparative Observer for Conserved Areas among Whole Genomes

SeungHeui Ryu,[1] Hwan-Gue Cho,[2] DoHoon Lee[3]

1 Introduction

To find a special region in biological sequence is one of the hot issues such as motif finding in a sequence, searching homology between sequences, and multiple alignment for finding homology among sequences [1, 2, 3]. Furthermore, identifying the function and relationship among the reserved region or genes has been interested like gene clustering, gene order, and gene pair ordering problem [4]. Sometimes the sequence between conserved two regions may give an information us to find mechanism in biological change. For example, these small size sequences between specific genes or conserved regions could be identified as a marker to express specific diseases among genomes [6]. For the purpose, many tools for small size comparison like BLAT [5] could be available. However these approaches have limit to extract the small gap between given regions. We intend to cave specific sequence and conserved sequence as a marker through finding homologies in a genome and analyzing similar sequence between genomes.

2 COCAW System

COCAW[4] (Comparative Observer for Conserved Areas among Whole Genomes) compares sequences between Genomes at start points as selected arbitrary two positions k-mer. Finding valuable region like marker is eventually our goal. Because the search space for finding homology pair using k-mer pattern is vary huge in a genome or among genomes, it is necessary to verify that there are meaningful pair regions in advance. COCAW is a system for the verifying and visualizing the pair region among genomes. The procedure is the followings: (1) two query positions are selected in a reference genome, (2) k-mer pattern is obtained respectively, (3) find k-mer pattern pair in a genome and extend given two patterns until increasing reasonable size, (4) extracting all short sequence,gap, between extended patterns, and (5) repeat step (3) and (4) among other genomes. COCAW has the following features: Web based service, display the conserved pair region, report sequences of all conserved pair region, extract and saving sequence of gaps, and link to NCBI sequence viewer.

In the figure 1, reference genome is a *Streptococcus agalactiae* (NC 004368) bacteria. Compared genomes are G1 and G2 respectively. G1 is a *Klebsiella pneumoniae* (NC 009648) bacteria and G2 is a *Pseudomonas aeruginosa* (NC 002516). Same color is a similar conserved region. Those figure show there are few conserved pair regions. Two numbers of pairs are estimated similar conserved regions after second position 254116 sequence in a reference genome. Six pairs and a pair are estimated similar conserved regions in G1 and G2 respectively.

Figure 2 shows the web interface of COCAW. We use 14 bacteria and 6-mer pattern database that have each position. It shows that COCAW can service multiple alignment of pair regions via multiple choice of genome.

3 Conclusion

COCAW is a tool for multiple alignment of given pair region and visualizing. From the our approach we can extract the short sequence between conserved regions and make an analysis of the short sequences for characterization of conserved region or genes. Furthermore we can characterize the short region such as disease marker.

[1]School of Computer Science and Engineering, Pusan National University, Pusan, Korea. Email: bogeum@pusan.ac.kr
[2]School of Computer Science and Engineering, Pusan National University, Pusan, Korea. Email: hgcho@pusan.ac.kr
[3]School of Computer Science and Engineering, Pusan National University, Pusan, Korea. Email: dohoon@pusan.ac.kr
[4]This work was supported by Blue Ocean Project of SMBA, Korea.

COCAW is a progressive tool, so additional functions are required: easy interface for handling, link to the other tools, and display useful information getting from Genbank and COG. COCAW is available on `http://164.125.37.216/Projects/COCAW`.

References

[1] Kenzie D, Maclassc, and Ernest Fraenkel, 2006. Practical Strategies for Discovering Regulatory DNA Sequence Motifs. *PLOS Computational Biology*, 2(4):201–210.

[2] Stephen F. Altschul, Warren Gish, and Webb Miller, 1990. Basic Local Alignment Search Tool. *J. Mol. Biol.*, 215(3):403–410.

[3] Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson, 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680.

[4] Ross Overbeek, Michael Fonstein, Mark DSouza, Gordon D. Pusch, and Natalia Maltsev, 1999. The use of gene clusters to infer functional coupling. *PNAS USA*, 96:2896–2901.

[5] W. James Kent, 2002. BLAT—the BLAST-Like Alignment Tool. *Genome Research*, 12:656–664.

[6] JunHyung Park, 2006. *Universal Probe Design Strategy for Mocrobe Identification*. Pusan National University.

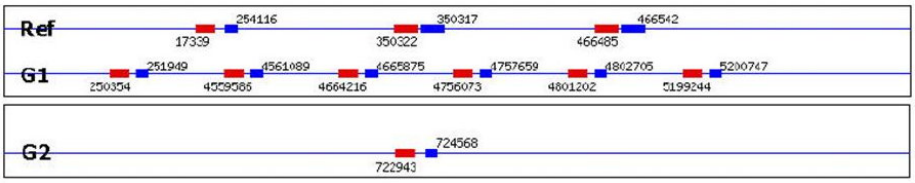


Figure 1: Reference genome is a *Streptococcus agalactiae*, G1 *Klebsiella pneumoniae*, and G2 *Pseudomonas aeruginosa*. In reference genome, the pair are found 3 times. In G1 and G2, 6 pairs and a pair are found respectively.

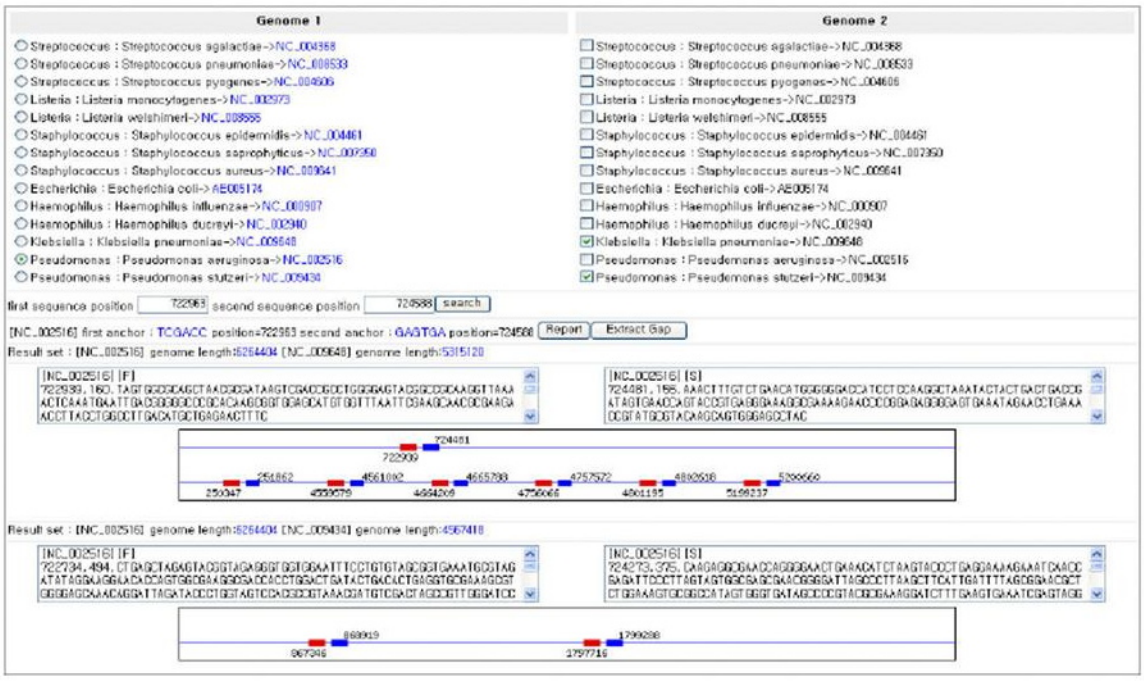


Figure 2: COCAW interface: Genome1(up left) is reference genome, genome2 are compared genomes, multiple genomes choice(up right). The next is a getting queries of position of 6-mer pattern, and then the sequences and positions are displayed.

SPIKE: Signaling Pathways Integrated Knowledge Engine

Ran Elkon,[1] Rita Vesterman,[1] Nira Amit,[1] Igor Ulitsky,[2] Gilad Mass,[1] Idan Zohar,[2] Dorit Sagir,[1] Jackie Assa,[2] Yosef Shiloh,[1] Ron Shamir[2]

Biological signaling pathways that govern cellular physiology form an intricate web of tightly regulated interlocking processes. Data on these regulatory networks are accumulating at an unprecedented pace. The assimilation, visualization and interpretation of these data have become a major challenge in biological research, and once met will greatly boost our ability to understand cell functioning on a systems level.

To cope with this challenge, we are developing the SPIKE knowledge-base of signaling pathways. SPIKE contains three main software components: 1) A database (DB) of biological signaling pathways. Carefully curated information from the literature and data from large public sources constitute distinct tiers of the DB. 2) A visualization package that allows interactive graphic representations of regulatory interactions stored in the DB and superposition of functional genomic and proteomic data on the maps. 3) An algorithmic inference engine that analyzes the networks for novel functional interplays between network components.

SPIKE implements user-friendly data submission forms which allow registered users to upload data to SPIKE DB. Our vision is that the DB will be populated by a distributed and highly collaborative effort undertaken by multiple groups in the research community, where each group contributes data in its field of expertise.

The integrated capabilities of SPIKE make it a powerful platform for the analysis of signaling networks and the integration of knowledge on such networks with omics data. SPIKE is available at `http://www.cs.tau.ac.il/ spike`.

Acknowledgments. The development of SPIKE was supported by the A-T Children's Project, by ESBIC-D, a coordinated action under the EU Sixth Framework, and by a Converging Technologies grant from the Israeli Science Foundation.

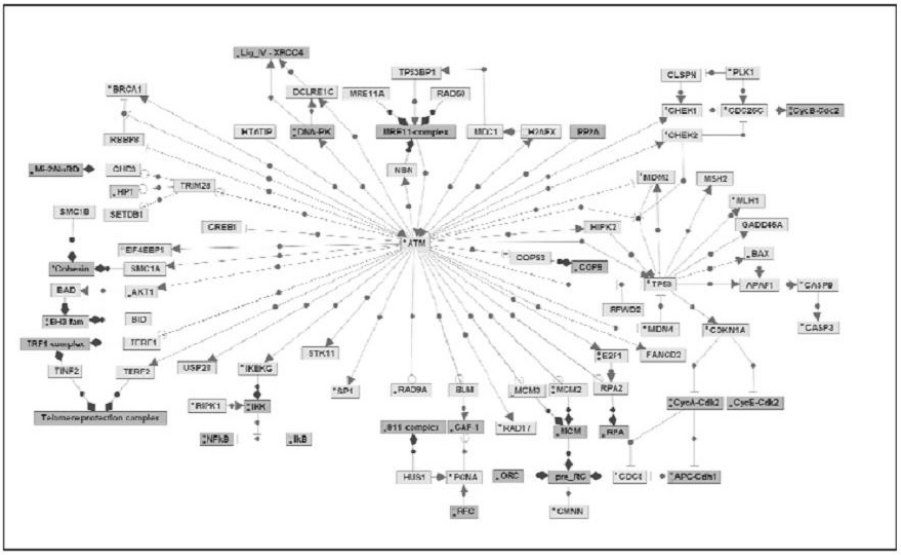


Figure 1: An example of SPIKE signaling map.

[1]The David and Inez Myers Laboratory for Genetic Research, Department of Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel.
[2]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel.

GPU-Based Fast K-means Clustering of Gene Expression Profiles

S. A. Arul Shalom,[1] Manoranjan Dash,[2] Minh Tue[3]

1 Introduction

K-means well suits gene expression profiles clustering, assuming that most gene clusters are spherical or elliptical in shape. Usually gene expressions have high dimensions say 4 to 300 experimental measurements, with a few thousands to hundreds of thousands of genes. K-means algorithm is to identify patterns in gene expression profiles in human mammary epithelial cells (HMEC) and breast cancers [2], and to analyze oligonucleotide microarray gene expression data (GDS958) [1]. Often it is time consuming while analyzing clusters from large microarray experiments. We present here how to harness programmable Graphics Processing Unit (GPU) to speed-up k-means cluster computations.

2 Computational Performance of GPU over the CPU

The application of GPU today has been well proven beyond the traditional processing of vertices and fragments [3]. One major arena in the application of GPGPU is the analysis of Gene expression profiles in various forms of micro array data. A phenomenal computation speed up to 4x to 12x can be achieved using todays commodity GPU when compared to the use of CPUs on high-end desktops. The factors that enable the processing power of GPUs are the inherent parallel architecture, high peak memory bandwidth, possible high floating-point operations, and the various stages of programmable processors which are shown in figure 1. We have demonstrated the performance gain of using GPU for k-means computation via experimentation. GPU performance (Time/Iteration) of k-means implementation is compared with the CPU implementation as shown in figures 2 and 3.

3 GPU Implementation of K-means Clustering

Micro array experiments produce massive amounts of data, which needs to be analyzed for significances in expression pattern. Often, computational methods such as k-means clustering need to be used for this purpose [1, 2]. In this section we brief the steps to implement k-means clustering of microarray data on the GPU efficiently. The two sets of microarray data: HMEC [2], and GDS958 [1] are used. The kernels are implemented in the fragment processor shown in figure 1 and executed via shaders. It is seen that the GPU implementation is faster than the CPU by 7x to 8x. The steps are as follows: Transfer gene expression data to GPU textures from CPU arrays. Store Gene expressions in textures using the "Luminance" format. Parallel computation of gene distances and minimum distances (figure 4). Parallel grouping of gene textures based on minimum distances. Update centroids in GPU and transfer of cluster information to CPU. A part of the shader implementation in OpenGL Shading Language (GLSL) for distance computation is given in figure 7.

4 Summary of Experiment Results

The micro array gene expression profiles have been represented via GPU textures and the k-means computations are executed via fragment shaders for data size over 1M. Our approach in clustering avoids the need for data and cluster information transfer between the GPU and CPU in between the iterations thus improving the efficiency 2x compared to a previous [4] GPU implementation of k-means clustering. The GPU k-means clustering implemented for the HMEC gene expression data set with about 65500 genes has shown tremendous speed gain of 7x to 8x against the CPU implementation. The efficiency gets higher as the sizes of the dimensions get larger. Gene clustering is done for the GDS958 Gene expression data set (16800 genes and 4 dimensions) and the 4 clusters are visualized in figure 5. It is evident that traditional clustering methods on large sets of micro array data can be done fast. Our future works will be on implementing hybrid k-means and hierarchical methods on the GPU.

[1]School of Computer Engineering, Nanyang Technological University, Singapore. Email: sall0001@ntu.edu.sg
[2]School of Computer Engineering, Nanyang Technological University, Singapore. Email: asmdash@ntu.edu.sg
[3]NUS High School, Singapore. Email: h0630082@nus.edu.sg

References

[1] De, K. R., Bhattacharya, A. 2006. Identification of Over and Under Expressed Genes Mediating Allergic Asthma. Springer: 943–952.

[2] Liang, J., Kachalo, S. 2002. Computational analysis of microarray gene expression profiles: Clustering, classification, and beyond. *Chemometrics and Intelligent Laboratory Systems*, 62:199–216.

[3] Owens, J. D., Luebke, D., Govindaraju, N., Harris, M., Krueger, J., Lefohn, A. E., Purcell, T. J. 2005. A Survey of General-Purpose Computation on Graphics Hardware. *Eurographics State of the Art Reports*, pages 21–51.

[4] Hall, J. D., Hart, J. C., 2004. GPU Iteration of Accelerated Clustering. In: *ACM Workshop on GPC on GPU at SIGRAPH 2004*.

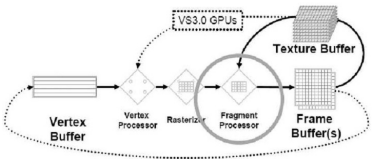


Figure 1: Programmable stages of a modern GPU.

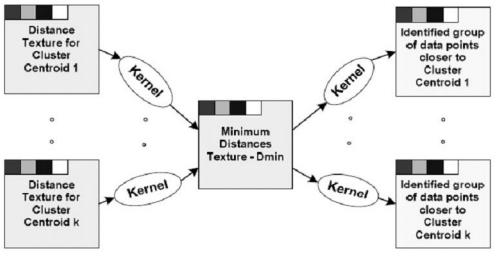


Figure 4: Identification of Cluster groups from the Distance textures.

No. of Clusters	Gen Expr Dimensions	Time in Sec/Iteration GPU 8600	CPU P4	Efficiency
3	4	0.00623	0.04954	8.0
4	4	0.00902	0.06521	7.2
3	9	0.01251	0.10311	8.2
4	9	0.01589	0.13331	8.4

Figure 6: HMEC gene expression clustering (65500 genes): Comparison between GPU & CPU performance.

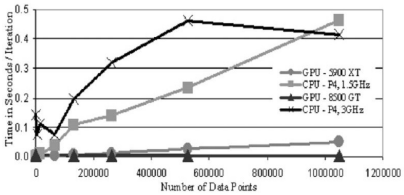


Figure 2: GPU NVIDIA 5900 & 8600 vs CPU P4 based on data size. GPU is 4x to 12x faster. The speed gain increases as the size of data sets get larger.

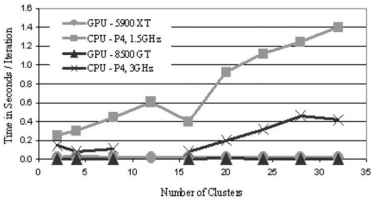


Figure 3: GPU NVIDIA 5900 & 8600 vs CPU P4 based on cluster size. GPU 5900 is 10x to 20x faster and 8600 is about 50x faster when there are more than 20 clusters.

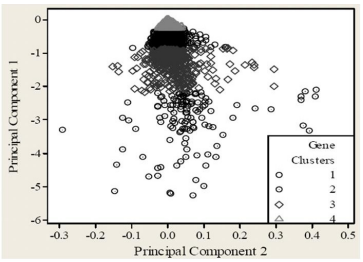


Figure 5: GDS958 Gene expression data clusters formed via k-means implementation on GPU.

```
char* shader7 = \
"#extension L_ARB_texture_rectangle : enable\n" \
"uniform sampler2DRect texture;" \
"void main(void) { " \
" float val1 = texture2DRect(texture,
gl_TexCoord[0].st).x;" \
" float val2 = texture2DRect(texture,
gl_TexCoord[1].st).x;" \
" float val3 = texture2DRect(texture,
gl_TexCoord[2].st).x;" \
" float val4 = texture2DRect(texture,
gl_TexCoord[3].st).x;" \
" gl_FragColor.x = (val1+val2+val3+val4);"\" "}";
```

Figure 7: GLSL codes for new centroid computation.

Stochastic Switching Behavior of a Bistable Auto-Phosphorylation Network

Marvin N. Steijaert,[1] Huub M.M. ten Eikelder,[1] Anthony M.L. Liekens,[1] Dragan Bosnacki,[1] Peter A.J. Hilbers[1]

1 Introduction

Cells utilize intricate protein reaction networks to sense, transfer and interpret information about their envi- ronments. Many of these networks are based on cycles of phosphorylation and dephosphorylation reactions [2]. Although ordinary dierential equations (ODEs) provide a useful tool to describe the interactions within those networks, they fail to take into account the molecular species for which only small numbers of particles are available. For these cases, more detailed stochastic methods are required. Here we consider a phosphorylation network with auto-phosphorylation. When modeled with ODEs, such a network can act as a bistable switch. We study to what extent such an auto-phosphorylation cycle with a small number of particles can still be considered a bistable switch. In particular, we analyze the relative stability of the switch states as a function of the number of particles. To this end, we adopt a Markov model approach.

2 Stochastic Models of Auto-Phosphorylation

The smallest observed phosphorylation module that can theoretically yield bistability is a single phosphorylation cycle with (trans-) auto-phosphorylation. In such a cycle the phosphorylated species catalyzes its own production from the dephosphorylated species. We observed that 113 out of 273 kinases (i.e., enzymes that catalyze the addition of phosphate groups to proteins) in the Human Protein Reference Database [3] show auto-phosphorylation.

Here we study such a module with an input kinase S as shown in Figure 1a. We describe all reactions with Michaelis-Menten kinetics and choose an appropriate set of dimensionless parameters: we keep the concentration of phosphatase (i.e., the enzyme that catalyzes the dephosphorylation reaction) at 1 and choose for all reactions the Michaelis constant $K_m = 0.05$ and the catalytic constant $k_{cat} = 1$, which leads to the ODE equilibria shown in Figure 1b.

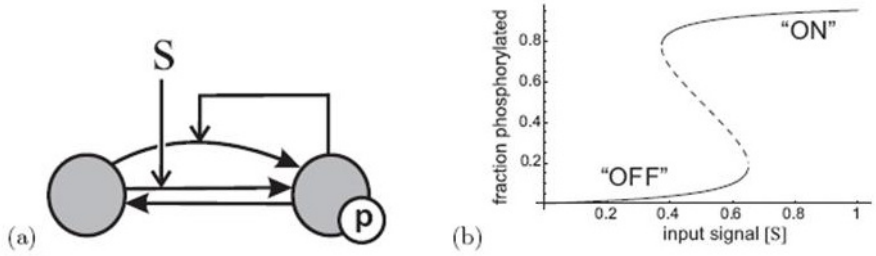


Figure 1: (a) Phosphorylation cycle with auto-phosphorylation and input kinase S. (b) For appropriate parameter sets, the ODE model displays bistability (solid lines: stable equilibria, dashed line: unstable equilibria).

For a finite number of particles N and input signal $[S]$, this network can also be described as a continuous time Markov model, which runs on the same time scale as the ODE model. From this description, we can directly derive the stationary distribution of the corresponding system. For systems with 10, 100 and 500 particles and a constant input concentration $[S] = 0.54$ (an arbitrarily chosen value within the bistable range), these distributions are shown in Figure 2a. In all these distributions there are two peaks (hereafter, the first and second maxima are referred to as "OFF-state" and "ON-state", respectively), which correspond with the stable equilibria of the ODE. The probabilities of intermediate

[1]Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. Email: m.n.steijaert@tue.nl

states are relatively large for small values of N, but decrease with increasing N. Note that for this value of $[S]$ the ratio between the probabilities of the OFF- and ON-states decreases with increasing N. This seems to imply that the OFF-state disappears for very large values of N. However, this is not the case. In fact, we have found that the ratio between the probabilities of the OFF-state and the intermediate minimum increases exponentially with growing N.

By analyzing the expected time for the system to move from one maximum to the other, we can show that for large N the OFF-state remains quasi stable. These transition times can be derived from the Markov model description [1, 4] and are shown in Figure 2b. As expected, it takes far less time to go from the OFF-state to the ON-state than it takes to switch in the opposite direction. However, both transition times grow exponentially fast with N, hence decreasing the probability of switching in either direction. Consequently, when N is suciently large, the behavior of the system in a reasonable long time is determined by the initial state of the system, yielding a so-called quasi-stationary distribution. In other words, for large values of N, the switch will remain in its initial state for a very long time. Of course, ultimately the information of the initial state is lost and the system is described by the stationary distribution.

So far, we have only considered one specific value of $[S]$. Yet, the above holds for any value of $[S]$ within the bistable range of the ODE in the sense that in the stationary distribution there are always two peaks between which the expected switching time grows exponentially fast with N. Note that for lower values of $[S]$ the OFF state is dominant in the stationary distribution, while for higher values of $[S]$ the ON state is dominant. Nonetheless, in either case both states are effectively stable for large enough values of N.

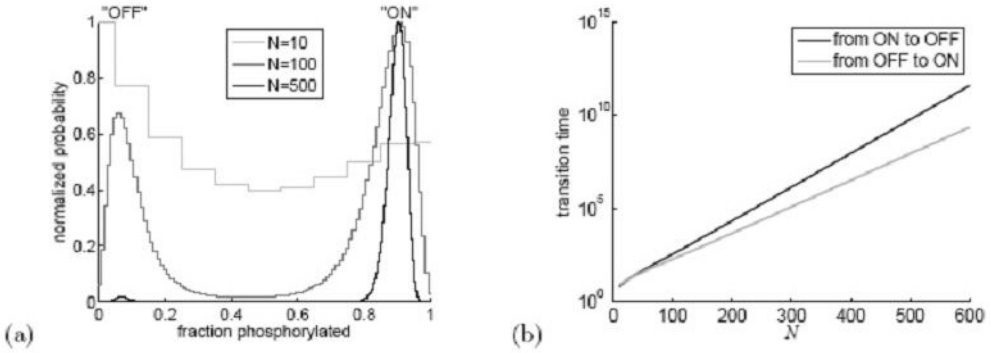


Figure 2: (a) Stationary distributions (normalized to a maximum value of 1) for $[S] = 0.54$ and $N = 10, 100, 500$. (b) Transition times between two peaks of stationary distribution for $[S] = 0.54$.

3 Conclusion

With the given parameters the ODE model of the auto-phosphorylation network has a stable OFF-state for small values of $[S]$ and a stable ON-state for large values of $[S]$. In the overlapping range, the system is bistable. For small values of N, the stochastic description shows clearly different behavior. As shown by the short transition times, the system can easily switch between both states and does therefore not act as a bistable switch. However, the switching time in both directions grows exponentially fast with N. Hence, as N grows to infinity, the stochastic model's behavior converges to that of the ODE model.

References

[1] Allen, L.J.S., 2003. *An Introduction to Stochastic Processes with Applications to Biology.* Upper Saddle River, NJ: Pearson.

[2] Cohen, P. 2000. The regulation of protein function by multisite phosphorylation—a 25 year update. *Trends Biochem Sci*, 25(12):596–601.

[3] Human Protein Reference Database, Release 7, `http://www.hprd.org`.

[4] Kampen, N.G. van, 1992. *Stochastic Processes in Physics and Chemistry.* Amsterdam, NL: Elsevier.

Ubiquitin-Proteasome Pathway Genes and Prostate Cancer

B Bharath Bhat, Ashwin Ram B, Arathi Raghunath, Khamari Lokanath, Sadanandan Vidyendra, Jignesh Bhate, Usha Mahadevan[1]

1 Introduction

Ubiquitin-proteasome system plays a significant role in pathogenesis, particularly in cancer [1]. Understanding degradation of proteins through this pathway is important in short listing targets and drug design. Here proteasome regulated genes in general and specifically those involved in prostate cancer are analyzed using NetPro^{TM}, a hand curated interactome knowledgebase. The analysis brings out the utility of the database to understand molecular interactions and also importance of ubiquitin-proteasome pathway in cancer.

2 Method and Discussion

"Ubiquitin pathway module" of NetPro^{TM}, was explored for proteins involved in ubiquitin-proteasome pathway using WebMINE, a Java based query interface. The query included interaction terms pertaining to ubiquitination, degradation, expression and also 6 proteasome inhibitors (Benzyloxycarbonylleucyl-leucyl-leucine aldehyde, Bowman-Birk inhibitor, Clasto-lactacystinbeta-lactone, epoxomicin, Lactacystin, PS 341).

(A) Regulation of ubiquitin-proteasome pathway involved genes. Table 1 summarizes results with combinations of queries and key processes associated with the group of genes as given in GO ontology. Interactions that lead to degradation or up regulation of the proteins by this pathway were considered by using appropriate interaction terms. NetPro^{TM} enabled to distinguish regulation mechanisms of ubiquitin-proteasome pathway interacting molecules. About 1/7th of the proteins annotated as ubiquitin ligase (this includes all evidences like IEA also) in GO have proven interaction with proteins that are degraded by proteasome pathway. 23 proteins degraded by proteasome pathway have no known ubiquitin ligases. It is possible that some of these genes do not go through ubiquitination for degradation by proteasome. Among the 213 genes that interact with ubiquitin ligases, $\sim$ 40 of the genes are involved in apoptosis, $\sim$ 25 in cell cycle by GO annotation, which in total is about 1/4th of the ubiquitin ligase interactors.

(B) Ubiquitin pathway and prostate cancer. Two sets of genes, genes involved in pathogenesis and genes differentially regulated by prostate cancer drug Taxotere were analyzed. Results are given in Table 2. Of the 31 genes associated with pathogenesis of prostate cancer, 19 were found to be targeted by ubiquitinproteasome pathway for degradation [2]. Overall 16 of the 32 genes influenced by Taxotere in PC-3 or LNCaP are involved in ubiquitin-proteasome pathway [3, 4].

To summarize, the results show predominant regulation of apoptosis, cell cycle or transcription processes linked genes by ubiquitin-proteasome pathway. 50% or more of genes involved in prostate cancer pathogenesis or drug action were found to be regulated by ubiquitin-proteasome pathway. Further, networks of proteins involved in the proteasome pathway could be generated to understand the regulation mechanisms involved (using a visualization tool).

References

[1] Nalepa, G., Rolfe, M. and Harper, J.W. 2006. Drug discovery in the ubiquitin-proteasome system. *Nat Rev Drug Discov*, 5(7):596–613.

[2] Foley, R., Hollywood, D., and Lawler, M. 2004. Molecular pathology of prostate cancer: The key to identifying new biomarkers of disease. *Endocr Relat Cancer*, 11(3):477–488.

[1]Molecular Connections Pvt. Ltd., Kandala Mansions, 2/2, Kariappa Road, Basavangudi, Bangalore-560004, India. Email: usha@molecularconnections.com

[3] Li, Y., Hussain, M., Sarkar, S.H., Eliason, J., Li, R. and Sarkar, FH. 2005. Gene expression profiling revealed novel mechanism of action of Taxotere and Furtulon in prostate cancer cells. *BMC Cancer*, 5:7.

[4] Li, Y., Li, X., Hussain, M. and Sarkar, F.H. 2004. Regulation of microtubule, apoptosis, and cell cyclerelated genes by taxotere in prostate cancer cells analyzed by microarray. *Neoplasia*, 6(2):158–167.

Protein groups	Numbers	Predominant processes (GO)
Proteins classified as ubiquitin ligases by GO (all evidence codes)	623	Ubiquitination
Unique ubiquitin ligases (as classified by GO) with interactions in NetPro^{TM}	85	Ubiquitination
Unique proteins interacting with ubiquitin ligases (as classified by GO)	533	ND
Unique proteins down regulated by ubiquitin ligases	213	Apoptosis $\sim$ 40; Cell cycle - 25; Transcription - 19
Unique proteins up regulated by ubiquitin ligases	37	ND
Proteins influenced by proteasome inhibitors	171	ND
Proteins having interactions both with proteasome inhibitors and ubiquitin ligases (as classified by GO) in NetPro^{TM}	22	Apoptosis $\sim$ 5; Cell cycle - 2; Transcription - 3
Proteins having interactions only with proteasome inhibitors and not ubiquitin ligases (as classified by GO) in NetPro^{TM}	23	Apoptosis $\sim$ 7;Cell cycle - 2; Signal transduction - 6

Table 1: Ubiquitin pathway molecules in NetPro^{TM} and the associated processes, as of December 2007.

Condition of gene profile	Number of Genes involved in ubiquitin-proteasome pathway
Prostate cancer pathogenesis [2]	19 of 31
Regulated in the same way in PC-3 and LNCaP cell lines by Taxotere [3, 4]	7 of 12
Not affected in one or the other cell lines by Taxotere [3, 4]	9 of 18
Opposite expression pattern upon Taxotere treatment in the 2 cell lines by Taxotere [3, 4]	0 of 2

Table 2: Prostate cancer related genes involved in ubiquitin-proteasome pathway.

Fast Approximate Hierarchical Clustering using Similarity Heuristics and Adaptation to Time Constraints

Meelis Kull,[1,2,3] Jaak Vilo[1,2,3]

1 Introduction

Agglomerative hierarchical clustering (AHC) is a common unsupervised data analysis technique used in several biological applications, most often probably for gene expression data. Standard AHC methods require that all pairwise distances between data objects were known. With ever increasing data sizes this quadratic complexity poses problems that cannot be overcome by simply waiting for faster computers. We propose an approximate AHC algorithm HappieClust that can be tuned to obey user-given time constraints. The key to the algorithm is to limit the number of pairwise distances calculated to a fraction of all possible distances which are chosen by using similarity heuristics.

2 Methods

HappieClust first calculates a fraction of all pairwise distances and then performs a modified hierarchical clustering procedure, where merging decisions are based on the known distances only. Careful implementation achieves linear running time in the number of used distances. The main concern here is the quality of the resulting clustering, which depends heavily on the subset of distances chosen. To mimic the best greedy choices of the AHC we introduce a similarity estimation heuristics that helps to rapidly find pairs of similar data objects. The heuristic is based on the observation that if two objects are close enough to each other then the distance to any third object from both of these is approximately the same. We turn this observation upside down and look for pairs of objects which are approximately at the same distance from several other objects (which we refer to as pivots). Figure 1 illustrates the experimental result that these pairs are more probably similar. Further experiments have confirmed that using such similarity heuristics significantly increases the quality of resulting clustering. Our similarity heuristic closely relates to several similarity search algorithms [5].

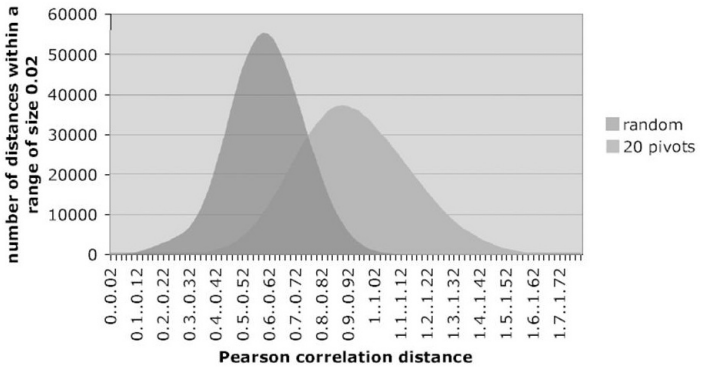


Figure 1: Distributions of 1000000 Pearson correlation distances in the data set by Shyamsundar et al [4]. The random choice of pairs results in a normal-like distribution with mean 0.9 whereas pairs chosen using similarity heuristics have mean 0.6.

[1]Institute of Computer Science, University of Tartu, Liivi 2, 50409 Tartu, ESTONIA. Email: `meelis.kull@ut.ee, jaak.vilo@ut.ee`
[2]Estonian Biocenter, Riia 23b, 51010 Tartu, ESTONIA.
[3]Quretec Ltd. Ulikooli 6a, 51003 Tartu, ESTONIA.

3 Quality Evaluation

Hierarchical clustering is often followed by a study to find subtrees with over-representation of genes annotated to some specific Gene Ontology term or pathway [1]. It would be desirable for the approximate hierarchical clustering to reveal mostly the same pathways and Gene Ontology terms. We used the web-based toolset g:Profiler [3] and looked for the highly over-represented terms (p-value less than g:SCS threshold [3]) in the subtrees of the full hierarchical clustering dendrogram. To evaluate approximate hierarchical clustering we looked how much these p-values differ in the logarithmic scale. Figure 2 shows an example of such evaluation where HappieClust was 25 times faster than the conventional full hierarchical clustering. The histogram contains p-value changes for 656 Gene Ontology terms and only 14 of these have dropped more than 50%, 139 have dropped 25–50%, 273 have dropped 0–25%, and 230 have become more significant. This indicates that we have found most of the biologically meaningful clusters 25 times faster.

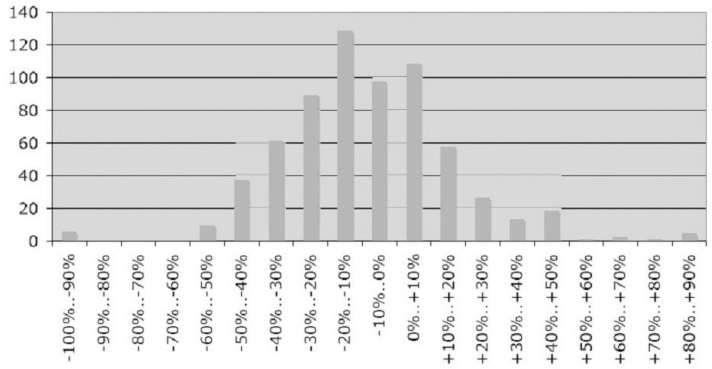


Figure 2: Gene Ontology and pathways annotations based quality for the data set by Lukk et al [2]. Histogram of changes in p-values from full clustering to HappieClust running 25 times faster.

4 Conclusions

Since approximate hierarchical clustering runs in near-linear time in the number of calculated distances, it is easy to estimate the running time in advance. This gives the great possibility to choose the number of distances to achieve the appropriate running time. Such feature is very useful in web-based applications where users expect fast response time.

The quality evaluation shows that dendrograms with useful information can be obtained several magnitudes faster than the conventional full hierarchical clustering.

Computational experiments verify that HappieClust is well suited for the large-scale gene expression analysis both on personal computers as well as public online web applications.

References

[1] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G. 2000. Gene ontology: Tool for the unification of biology. *Nat. Genet.*, 25(1):25–29.

[2] Lukk, M., Nikkila, J., Ukkonen, E., Brazma, A. 2007. Application of text mining to create human gene expression atlas from public data. Submitted.

[3] Reimand, J., Kull, M., Peterson, H., Hansen, J., Vilo, J. 2007. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*

[4] Shyamsundar, R., Kim, Y.H., Higgins, J.P., Montgomery, K., Jorden, M., Sethuraman, A., van de Rijn, M., Botstein, D., Brown, P.O., Pollack, J.R. 2005. A DNA microarray survey of gene expression in normal human tissues. *Genome Biology*, 6(3).

[5] Zezula, P., Amato, G., Dohnal, V., Batko, M. 2006. Similarity Search: The Metric Space Approach. Springer.

A Training-Set-Free Stochastic Model for Peptide Identification

Gongjin Dong, Yantao Qiao, Yu Lin, Shiwei Sun, Chungong Yu, Dongbo Bu[1]

1 Introduction

How to distinguish the exact peptide-spectrum matches from the random matches remains a challenge to peptide identification through tandem mass spectrometry. Various models and algorithms have already been proposed to solve this problem. Most of these models follows the machine-learning framework, and thus need a training set to derive the most likely parameters of the model. Unfortunately, preparing a training set is a tough requirement due to the heavy efforts to manually check thousands of peptide-spectrum pairs. This task becomes especially infeasible for the post-translation-modification (PTM) case: there are few training set for peptides with PTM since we know little rules about the fragmentation of peptides with PTM. In our previous work [1], we have proposed a stochastic model for peptide fragmentation. Though the model is effective, it still suffers from the requirement of a training set as input. Therefore, it is interesting and useful to design a training-set-free method for peptide identification.

We present in this paper such a training-set-free approach for peptide identification. As its name implies, the input must not be a training set with exact peptide-spectrum pairs; instead, the input can be a mixed one, i.e., a mixture of exact peptide-spectrum pairs along with random pairs. Though we do not know in advance whether a pair is exact or random, we can apply expectation-maximization (EM) technique to assign each pair a label automatically. The underlied assumption of this strategy is: (i) the exact peptide-spectrum pairs share a common peptide fragmentation pattern; (ii) however, we cannot derive a meaningful fragmentation model if using the random pairs as input. Analogously, the previous models follows a classification framework; while our method adopts a clustering framework.

We applied this approach to identify the false-positive pairs in SEQUESTs results. Experimental results suggest that this method can effectively detect the false-positive pairs reported by SEQUEST. Moreover, we performed database searching using this method and found that the performance is comparable with SEQUEST and MASCOT. By applying the method in this paper, we can enhance our previous work [1] to deal with spectrum with PTM. As result, we explored the fragmentation pattern of a peptide with PTM, and compared this pattern with the peptides without PTM.

We have implemented this method into an open source package, PI (Peptide Identifier), which can be freely downloaded from `http://www.bioinfo.org.cn/MSMS`.

2 Methods

The exact pairs and random pairs show significantly different similarity distributions (See Figure 1): the similarity score of exact pairs shows a log-normal distribution; while the similarity score of the random pairs shows a power distribution. For each pair in the input set, we run the following EM method to determine which distribution it lies in (See Figure 3). After the EM step, each peptide-spectrum was assigned with a probability that it comes from the score distribution of true positive or not. The peptide-spectrum pairs with high probability will be reported as true-positive results.

3 Results

3.1 Experiment 1: performing database search on Kellers data

First, we run SEQUEST and get top 500 candidate peptides for each spectrum. Then we apply our EM method to assign a probability for each possible pair. And this probability is used in the next training step as a weight. And these steps are iteratively carried out until the true positive score distribution is stable. Finally, our method reports 1481 pairs as true-positive. In contrast, SEQUEST reports 1486 pairs in this data set [2]. These results suggest that our method is comparable with SEQUEST.

[1]Bioinformatics Lab, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China. Email: {donggongjin, qiaoyantao, dwsun, bdb}@ict.ac.cn, biolinyu@gmail.com, yu_cg@hotmail.com

3.2 Experiment 2: apply iteration method on phosphopeptide data

Post Translational Modification (PTM) is extremely important because they may alter physical and chemical properties. John Yates et al have developed two statistical strategies to automatically evaluate serine/threonine phosphopeptide identifications on the basis of spectral features. However, these methods also suffer from the lacking a benchmark data set. Here, we applied our training set-free method to identify peptide through the spectrum with PTM: our method can identify 1394 pairs from a total of 1803 high-possible false-positive spectra [3]. The fragmentation pattern of a peptide with PTM is shown in Figure 2. This figure suggests: (i) phosphoserine(s) has similar fragmentation preference as serine(S); (ii) phosphothreonine(t) shows similar pattern to threonine(T), too; (iii) however, phosphotyrosine(y) differs a lot from tyrosine(Y). Specifically, the peptide bond L-y is more likely to break than L-Y, and V-y rarely to break compared with its counterpart V-Y. More specifically, the peptide bond y-P is rarely to break compared with s-P and t-P. This results suggest that a peptide with PTM may influence peptide fragmentation and generate a different spectrum from its original version without PTM.

References

[1] Yu CQ, Lin Y, Sun SW, Cai JJ, Zhang JF, Bu DB, Zhang Z, Chen RS. *J Bioinformatics and Computational Biology*, 2007.

[2] Keller A, Purvine S, Nesvizhskii AI, Stolyar S, Goodlett DR, Kolker E. *Omics*, 6:207–212, 2002.

[3] Bernd Bodenmiller, Johan Malmstrom, Bertran Gerrits, David Campbell, Henry Lam, Alexander Schmidt, Oliver Rinner, Lukas N Mueller, Paul T Shannon, Patrick G Pedrioli, Christian Panse, Hoo-Keun Lee, Ralph Schlapbach, and Ruedi Aebersold. *Molecular Systems Biology*, 3:139, 2007.

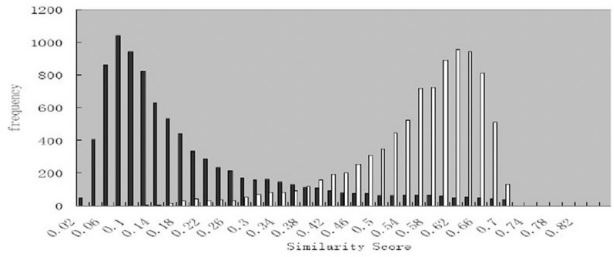


Figure 1: The Similarity Score Distributions for Exact Matches (left) and Random Matches (right).

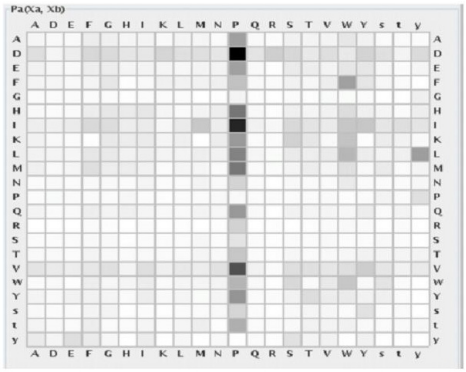


Figure 2: The cleavage preference for each peptide bond.

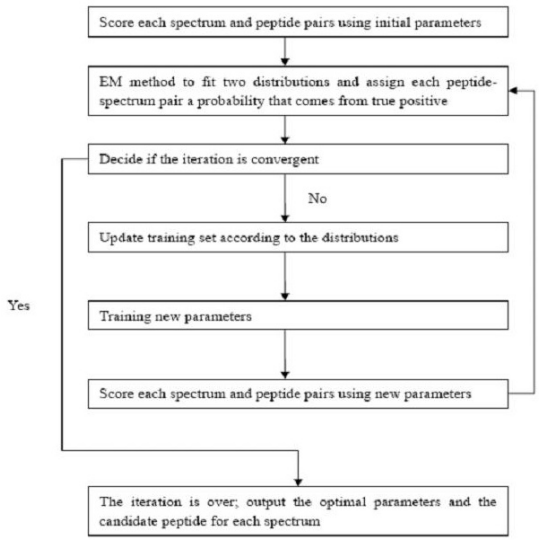


Figure 3: The peptide identifying procedure.

In Silico Charactization of Peptide Epitopes Recognized by Autoantibodies Present in IVIG Sample Preparations

Hans-Juergen Thiesen,[1] Peter Lorenz,[2] Zilliang Qian,[3] Yixue Li,[4] Michael Kreutzer,[5] Michael O. Glocker[6]

1 Introduction

Autoimmune diseases are diagnosed based on the presence of specific autoantibodies. Usually, the human organism generates antibodies for protecting the organism from invading viruses, bacteria, fungi and parasites. In case of proteins, these antibodies bind to conformational as well as linear epitopes. In general, the human body is genetically capable of generating a repertoire of more than 1012 different antibodies (immunoglobulins). Some of these antibodies called autoantibodies react with human tissues and molecules thereof. Most diagnostic techniques in autoimmune research rely on recombinant antigens on solid matrices (e.g. ELISA or line immunoassays) or on cell and tissue preparations (indirect immunofluorescence assays). Most assays use recombinant or purified autoantigens that are difficult to obtain and require great efforts of quality control. In contrast, the use of peptides would be more robust in that peptides can be easily synthesized in a standardized, reproducible and cost-effective manner, they can be covalently and specifically attached including spacers or any chemical modification and the microarrays are quite stable. Eventually, the current autoimmune diagnostics could be substantially replaced by the use of linear peptide epitopes. Hereto, a standardized workflow has been established how to identify linear epitopes, how to establish their binding characteristics and how to determine epitopes with higher binding affinities. Lastly, this information is expected to be essential for characterizing autoantibody subtypes (VH gene families), finally, leading to a more comprehensive understanding of autoimmune processes of humeral immune responses on the way to peptide(amino acid)-based digitized autoimmune diagnostics.

2 Material and Methods

Custom-made Replitope^{TM} high density peptide microarrays (JPT Jerini Peptide Technologies GmbH, Berlin, Germany) were used to profile epitope reactivities of intravenous immunoglobulin (IVIg) preparations. 15mer peptides originally synthesized by SPOT technology were spotted in triplicate divided as 3 subarrays onto activated standard-format glass slides. Staining involved blocking by serum albumin, staining with primary antibody and staining with fluorescently tagged secondary antibodies, each with extensive washing after each incubation. 16-bit images were recorded (Molecular Devices Axon Instruments GenePix 4000B) and spot intensities were evaluated against local background using Axon GenePix Pro software. Primary screens were performed on peptide microarrays containing 15mer overlapping peptides derived from major nuclear autoantigens relevant for diagnostics of systemic lupus erythematosus and systemic sclerosis. Secondary screens were done on microarrays that carried the top hit peptide of the first screen and systematic mutations thereof. Hereto, each position of the 15mer peptide was replaced by all remaining naturally occurring 19 amino acids. These microarrays carrying in total 300 peptides were incubated with affinity purified IVIg antibody species using columns made of cellulose material having attached the top hit peptide. Elution of bound antibodies with glycin pH 2.7 buffer resulted in the top hit peptide reactive fraction. In order to investigate whether particular rules of peptide antibody interactions can be derived from these data sets, affinity matrices were determined for each peptide po-

[1]Institute of Immunology, University of Rostock, Germany. Email: hans-juergen.thiesen@med.uni-rostock.de
[2]Institute of Immunology, University of Rostock, Germany. Email: peter.lorenz@med.uni-rostock.de
[3]Shanghai Institutes for Biological Sciences (SIBS), Chinese Academy of Sciences. Email: zl_qian@yahoo.com.cn
[4]Shanghai Institutes for Biological Sciences (SIBS), Chinese Academy of Sciences. Email: yxli@sibs.ac.cn
[5]Institute of Immunology, University of Rostock, Germany. Email: michael.kreutzer@med.uni-rostock.de
[6]Proteome Center Rostock, University of Rostock, Germany. Email: michael.glocker@med.uni-rostock.de

sition P1 to P15 and correlated with 544 matrices found in the Amino Acid Index Database (AAindex, `www.genome.jp/aaindex`). Pearson correlations were used to select the most informative matrices and heat maps were established thereof. Finally, this data analysis is used to predict putative amino acids characterizing the immunoglobulin binding to the peptides by making use of the Amino Acid Contact Potential Matrix Database, see AAindex as well.

3 Figures and Tables

High number of anti-epitope reactivities of natural antibodies in healthy antibody preparations: The analysis of linear epitopes using purified immunoglobulin preparations (IVIG) demonstrates that many of theses antibodies bind to linear peptides derived from human autoantigens as studied in the EU consortium Autorome (`www.Autorome.de`). To test the "background" of natural antibodies in the healthy population, three commercial preparations of IVIg (intravenous immunoglobulin; pool of IgG of > 10000 healthy donors) were profiled on peptide microarrays. All preparations that were prepared from different local blood donor populations (Israel, France, Italy) showed a sizeable number of specific epitope reactivities. Interestingly, a high proportion of these reactivities were shared between the different IVIg preparations suggesting that the natural antibody repertoire in the healthy Caucasian population encompasses considerable conservations (see Figure 1).

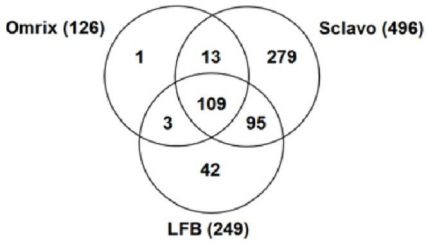


Figure 1: Venn diagram indicating the overlapping peptide reactivities of three commercial IVIg preparations at 0.5mg/ml IgG. Given are the number of reactive peptides in the different groups. Pearson correlations between the IVIg samples were between 0.62 and 0.79.

One of the highest epitope reactivities present in IVIg (15-mer mother peptide) was mutagenized at each position by generating peptides that carry single amino acids substitutions of 20 amino acids at each of the 15 positions of the peptide. In total, 300 mutants in triplicate were synthesized to determine the sequence specificity space of IVIg derived affinity purified antibodies. This peptide array was subjected to IVIg antibodies sample preparations that were purified by binding to the initial 15-mer mother sequence. Finally, binding matrices were determined qualitatively and quantitatively by ranking the binding scores of the substituted amino acids in peptide positions P1 to P15. This data set was studied to select peptides by higher binding scores compared to the initial mother sequence. Positions P4, P6, P10, P11, and P14 turned out to serve as dominant contact sites involved in antibody-peptide binding. Matrices describing physicochemical parameters (544 amino acid indices, see AAindex) were correlated with the binding preferences experimentally obtained. Finally, this data analysis is used to predict putative amino acids that might serve as contact sides on the surface of the antibody binding pouch. This data analysis might imply as well that a specific immunoglobulin subtype (VH gene family) might preferentially bind to these high affinity peptides. To validate whether the antibodies involved belong to one specific subclass of immunoglobulins, 2-DE gel electrophoresis followed mass spectrometric analysis are currently employed. Eventually, these data will be compared to known 3D-antibody-peptide structures.

Acknowledgments. We thank the EU for funding the Autorome project (`www.autorome.de`) (LSHM-CT-2004-005264).

References

[1] Lorenz, P., Kreutzer, M., Zerweck, J., Schutkowski, M., and Thiesen, H.-J. (2008). Probing the epitope signatures of IgG antibodies in human serum from patients with autoimmune disease. In: *Methods in Molecular Biology*, Schutkowski, M. et al. , editors, Totowa, NJ, USA, Humana Press Inc, Chapter 20b, in press.

Ultra-Deep Sequencing of Genetically Heterogeneous Samples

Niko Beerenwinkel,[1] Nicholas Eriksson,[2]
Volker Roth,[3] Osvaldo Zagordi[4]

1 Introduction

Ultra-deep sequencing is a family of new methodologies that, unlike traditional Sanger sequencing, typically give many short error-prone reads. They have emerged in the last few years and have been used for de novo sequencing, resequencing, genotyping, and sequencing of diseased genes [2]. While ultra-deep sequencing is now a commercially available technology for sequencing genetically homogeneous samples, the possibility to use it as a tool to estimate the population variation in a heterogeneous sample is a subject of active research. Biomedical applications range from intra-host virus populations [1, 3] to cancer cells derived from tumours.

2 Computational Approach

We propose a methodology to infer the different genomes present in the population (haplotypes) and to estimate their frequencies, which consists in the following four steps [1]:

1. **Alignment:** the existence of a reference genome to which the set of reads can be aligned is assumed. In order to correctly position the reads a pairwise alignment between each of them and the reference is performed taking the distribution of read errors into account;

2. **Error correction:** As specified above, the technique is characterized by a higher error rate and a shorter length of the reads. Nevertheless, its high coverage (many reads coming from the same region) allows an accurate reconstruction of the sequence. In this step we aim at correcting the errors of the sequencing process gaining information from multiple reads. In order to do so we have to distinguish technical errors from biologically relevant mutations;

3. **Haplotype reconstruction:** Once the technical errors have been corrected one is provided with many error free reads, each of them coming from a single unknown haplotype. The goal of this step is to reconstruct the smallest pool of haplotypes consistent with the observations;

4. **Haplotype frequency estimation:** Finally, assuming that the reconstructed haplotypes are extracted randomly from the population, one has to infer the population structure, i.e. the probability distribution on the set of haplotypes.

3 Error Correction via Local Clustering

We want to address here the error correction step. The key point is how to consider the differences between the reads and the reference sequence. In fact there are two possible sources of variance: technical errors due to the sample preparation and the sequencing process, and biologically relevant mutations distinguishing the observed haplotype from the reference one. The procedure adopted in [1] for local error correction consists in considering a window of overlapping reads and counting the number of mutations on single columns and pairs. If these are over-represented according to a statistical test, performed under the hypothesis of a single haplotype and a uniform error rate, a new haplotype is considered to be

[1]ETH Zürich Department of Biosystems Science and Engineering, Basel, Switzerland.
Email: niko.beerenwinkel@bsse.ethz.ch
[2]Department of Statistics, University of Chicago, IL, USA. Email: eriksson@galton.uchicago.edu
[3]Departement Informatik, University of Basel, Switzerland. Email: volker.roth@unibas.ch
[4]ETH Zürich Department of Biosystems Science and Engineering, Basel, Switzerland.
Email: osvaldo.zagordi@bsse.ethz.ch

present. Afterwards, a K-means clustering on the reads is run, with K equal to the number of haplotypes estimated. Once a satisfying partition of the reads is found, the consensus sequence is computed from the reads in a single cluster weighted by their quality scores.

4 Local Clustering via Dirichlet Process

A different technique pursued by us to achieve error correction consists in clustering reads applying a probabilistic Dirichlet process mixture (DPM), in a similar way to the method presented in [4] for haplotype inference in diploid populations. This approach allows to model a set of data points as coming from a set of equivalence classes, each with its prior distribution indexed by a parameter ϕ. The probability that assigns to each class a value of ϕ (or, equivalently, a distribution) is called the base measure of the process. In our case the base measure is a joint distribution on the pool of actually present haplotypes and on the pattern of technical reading errors. As a simple case we consider a uniform distribution on the haplotypes and, independent of it, a uniform error rate on the reads. The conceptual advantage of such a process is to define a probability distribution starting from the basic features of the system under investigation, rather than applying a general purpose clustering procedure. Moreover, as the only parameter governing the DPM also controls the generation of new clusters (haplotypes), we have access to a transparent estimate of the complexity of the model.

5 Application

We are testing the technique described above on both simulated and real data. In particular we compare the DPM error correction with the K-means based one, and we also compare it together with the haplotype reconstruction and frequency estimation steps. In order to perform controlled experiments we simulate the sequencing error process by means of the dedicated software ReadSim applied to a population derived from a HIV sequence. We also analyse read sets obtained from sequencing genetically diverse HIV samples. These viruses populations have been derived from several infected patients under antiretroviral therapy.

References

[1] Nicholas Eriksson, Lior Pachter, Yumi Mitsuya, Soo-Yon Rhee, Chunlin Wang, Baback Gharizadeh, Mostafa Ronaghi, Robert W. Shafer, Niko Beerenwinkel. Viral population estimation using pyrosequencing. `arXiv.org:0707.0114`.

[2] Marcel Margulies et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380, 2005.

[3] Chunlin Wang, Yumi Mitsuya, Baback Gharizadeh, Mostafa Ronaghi, Robert W. Shafer. Characterization of mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance. *Genome Research*, 17:1195–1201, 2007.

[4] Eric P Xing, Michael I Jordan, Roded Sharan. Bayesian haplotype inference via the Dirichlet process. *J Comput Biol*, 14(3):267–284, 2007.

Microarray Gene Recognition Using Multiobjetive Evolutionary Techniques

Dong L. Tong,[1] Robert Mintram[2]

1 Introduction

In this research we present a new simple and effective algorithm namely Genetic Algorithm Neural Network (GANN), to identify the underlying genes for a specific cancer classification using microarray data. The proposed algorithm exploits the multiobjective evolutionary capability of Genetic Algorithms (GAs) with the universal computational power of Multilayer Perceptrons (MLPs). In brief the GAs are used to identify relevant genes within the context of a classification task whilst the MLPs are applied as the cancer classifiers. The philosophy of the algorithm is to use the parameter settings that are no more complex than that required for the solution to the problem. The main objective is to analyse the implications of activation function, population size and fitness evaluation on the identification of relevant genes in discriminating microarray data.

With the use of microarray technology, the cancer classification accuracy has tremendously improved and benefits cancer diagnosis and prognosis. However, the identification of important genes is still in debate as many existing classifiers had been claimed better than the others. Further, high dimensional of irrelevant genes in microarray data, due to the presence of noise in DNA samples and noise produced during microarray processing, and the heterogeneity gene combination solutions to the cancer diseases had complicated the gene identification problem.

Thus, we propose a simple and alternative way to interpret microarray gene expression data. The novelty of this algorithm is the use of new fitness functions to derive the classification process. The acute leukaemia dataset [1] which comprised of 2 leukaemia subclasses (ALL and AML), containing 72 samples that were hybridised to high-density oligonucleotide microarrays and consisted of 7129 genes, was used to evaluate the performance of the proposed algorithm.

2 The Genetic Algorithm Neural Network (GANN) Approach

The basic paradigms of Genetic Algorithms (GAs) and Multilayer Perceptrons (MLPs) with minimal parameter settings are used, followed The Occam's Principle. The GA parameter settings are tournament selection size of 2, single-point crossover, mutation rate of 0.1, population size of 100, 200, and 300, and fitness evaluations of 5000, 10000, 15000 and 20000. A simple 3-layer feedforward neural network (FNN) with the structure of 10:5:2 (i.e. 10 input neurons, 5 hidden neurons and 2 output neurons) was implemented to compute the fitness of the genes with 8 different activation functions (binary, binary sigmoid, bipolar, bipolar sigmoid, integer, linear, tanh and threshold). We experimented all combinations of population size, fitness evaluation and activation functions. Further, we also performed the experiments with similar parameter settings for pre-processed leukaemia dataset with the data values in range of $[0, 1]$, according to the correlation between between genes.

The GANN method is described in brief as follows: First, a set of parameters were initialised before the beginning evolution of GANN model, including the operating parameters of GA and FNN. Second, the training begins by generating an initial random population of solutions (i.e. Chromosomes). Third, random select a chromosome and compute the fitness score of the chromosome using FNN. This continues until all chromosomes in the population have assigned the fitness scores. Fourth, 2 chromosomes were selected using tournament selection for reproduction and an offspring is produced. Fifth, the fitness of the offspring is evaluated using FNN. Sixth, the worst chromosome in the current population is replaced by the offspring and the entire chromosomes in the population is copied into the new generation (population). The new generation, which has equal size to the original population, begins the cycle again. This continues until the termination criterion (i.e. fitness evaluation) is met and the GANN stops. Seventh, the number

[1]Department of Design, Engineering and Computing, Bournemouth University, UK. Email: dltong@bournemouth.ac.uk
[2]Department of Design, Engineering and Computing, Bournemouth University, UK. Email: rmintram@bournemouth.ac.uk

of corrected classification sample and the frequency of genes selected for the classification are recorded. Steps 2 to 7 are performed 5000 times with the same parameter settings. Since we only interested in identifying the significant genes for classification instead improving classification accuracy, the entire dataset is used for training.

3 Experiment Results

Tables 1 and 2 show the best results obtained by GANN using 8 different activation functions. The GANN had correctly classified the average of 71.94 out of 72 samples on the original dataset and these classifications were achieved in 4 types of activation functions, which is binary function, bipolar function, bipolar sigmoid function and tanh function. While the best classification accuracy for the normalised dataset is slightly lower than the original dataset with the average of 71.93 achieved using linear activation function. The number of genes involved in the classification has been significantly reduced when we the dataset is pre-processed before the classification to be performed.

Activation Function	Original Dataset	Pre-processed Dataset
Binary	71.94* (P300.E20000)^o	71.44 (P300.E20000)
Binary sigmoid	71.93 (P300.E20000)	71.89 (P300.E20000)
Bipolar	71.94 (P300.E20000)	71.40 (P200.E20000)
Bipolar sigmoid	71.94 (P300.E20000)	71.90 (P300.E20000)
Integer	71.77 (P300.E20000)	71.86 (P300.E20000)
Linear	71.79 (P300.E20000)	71.93 (P300.E20000)
Tanh	71.94 (P300.E20000)	71.85 (P300.E20000)
Threshold	71.62 (P300.E20000)	71.91 (P300.E20000)

Table 1: GANN: The comparison result on the best average classification accuracy.

Activation Function	Original Dataset	Pre-processed Dataset
Binary	6904 (P300.E20000)	6898 (P300.E20000)
Binary sigmoid	6899 (P300.E5000)	6836 (P200.E20000)
Bipolar	6892 (P200.E15000)	6887 (P200.E20000)
Bipolar sigmoid	6887 (P300.E20000)	6834 (P300.E15000)
Integer	6950 (P300.E20000)	6780 (P300.E15000)
Linear	6946 (P300.E15000)	6794 (P300.E20000)
Tanh	6892 (P300.E10000)	6874 (P300.E20000)
Threshold	6969 (P300.E15000)	6805 (P300.E10000)

Table 2: GANN: The comparison result on the minimum number of genes involved in the classification.

* - The number of correctly classified samples out of 72 samples.

o - The indication of the parameter settings used. For example, P300.E20000 indicates population size of 300 and fitness evaluations of 20000.

References

[1] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–536.

A Heuristic Clustering Algorithm Using Graph Transitivity

Marcel Martin,[1] Sven Rahmann[2]

1 Introduction

We have developed a clustering algorithm based on *graph transitivity*.

Let $G = (V, E)$ be an undirected graph. V is a set of objects and $\{u, v\} \in E$ if u and v are "similar". The precise meaning of "similar" depends on the application. In our case, the objects are proteins, which are defined to be similar if their sequence similarity, as measured by the $-\log$ BLAST-E-value, exceeds a given threshold.

In the following, we write uv to denote the edge $\{u, v\}$.

An unweighted graph $G = (V, E)$ is transitive if and only if the following holds for all u, v, $w \in V$:

$$uv \in E \text{ and } vw \in E \Rightarrow\ uw \in E \tag{1}$$

This implies that a transitive graph is a disjoint union of cliques (see right side of Fig. 1).

A graph obtained from real-world data is hopefully "almost transitive". The problem is to make it transitive by using as few edge deletions and additions as possible. This can be seen as removing false positives and re-adding false negatives in experimental data. We call the resulting graph a *best transitive approximation*.

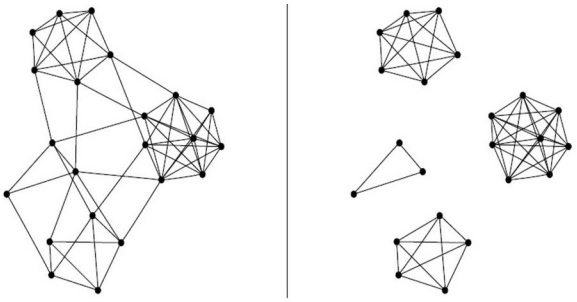


Figure 1: Left: original graph. Right: a transitive approximation.

Problem 1 (Transitive Graph Projection) *Given $G = (V, E)$, find a transitive $G^* = (V, E^*)$ such that $|E \triangle E^*|$ is minimal, where $\triangle$ is the symmetric set difference.*

This problem, also known as the *Cluster Editing Problem* or approximation of binary symmetric relations by equivalence classes [2], is NP-hard [1]. We developed a heuristic algorithm that gives approximate solutions.

We call triples (u, v, w) for which the implication of Eq. (1) does not hold conflict triples. That is, for which $uv \in E$, $vw \in E$, but $uw \notin E$. A graph is transitive if and only if it does not contain a conflict triple. Intuitively, the conflict triples are the defects in the graph that prevent it from being transitive.

Definition 1 (Deviation from Transitivity) *Define $D(G)$, the* deviation from transitivity *of a graph G, to be the number of its conflict triples.*

Definition 2 (Transitivity Improvement) *Let uv be an edge in the graph$G = (V, E)$. Removing it results in a graph $G' = (V, E \setminus \{uv\})$ with a possibly different number of conflict triples. We call $\triangle_{uv}(G) := D(G) - D(G')$ the* transitivity improvement *of edge uv.*

The *transitive closure* of a graph G is the minimum transitive graph containing all edges of G. In other words, this is a version of the problem in which only edge additions are allowed.

[1]TU Dortmund, Germany. Email: marcel.martin@tu-dortmund.de. Work done at Bielefeld University.
[2]TU Dortmund, Germany. Email: sven.rahmann@tu-dortmund.de

2 Algorithms

We observe: a) G is transitive iff $D(G) = 0$. b) Connected components of the graph can be treated separately. c) Only edge deletions need to be found (edge additions can be inferred by the transitive closure).

This results in the following idea for an algorithm: In each step, greedily remove the edge that yields the greatest transitivity improvement. When the graph gets split into two connected components, work on them recursively.

The procedure **RemoveCulprit**(G) removes the highest-scoring edge $\text{argmax}_{uv \in E}\{\triangle_{uv}(G)\}$ from G. The procedure $CC(G)$ assumes G is connected; it returns the total cost of all edge additions required for a transitive closure of G, $|V|(|V|-1)/2 - |E|$. $GH(G)$ returns the cost of all edit operations. It greedily removes edges until the graph consists of two connected components, for which it then solves the problem recursively. In detail, it works as follows:

1. $clcost \leftarrow$ **ClosureCost**(G)
2. Base case: If $clcost = 0$, return 0.
3. Set $delcost \leftarrow 0$, and repeat the following step until G consists of two connected components G_1 and G_2: **RemoveCulprit**(G) and $delcost \leftarrow delcost + 1$
4. Adjust *delcost* such that only deletions that contribute to the cut between G_1 and G_2 are considered, and re-add incorrectly deleted edges to G_1 and G_2.
5. Solve the problem recursively for G_1 and G_2 as long as there is a chance for a better solution:
   If $delcost \geq clcost$, return *clcost*.
   $cost_1 \leftarrow$ **GreedyHeuristic**(G_1)
   If $delcost + cost_1 \geq clcost$, return *clcost*.
   $cost_2 \leftarrow$ **GreedyHeuristic**(G_2)
   If $delcost + cost_1 + cost_2 \geq clcost$, return *clcost*.
6. Return $delcost + cost_1 + cost_2$.

3 Guarantees and Results

Input graphs resulting from actual experiments should be almost transitive since it makes no sense to find clusters in random graphs. If the "perturbations" (equivalent to measurement errors) are limited in a certain way, *we can guarantee that our algorithm finds the unperturbed input graph.*

Modification rule. Given a transitive graph T with clusters C_i, a vertex u in cluster C_i may get at most $\frac{2}{9}n_i$ additional edges and at most $\frac{2}{9}n_i$ of its edges may be removed.

Theorem 1 *If a graph G has been obtained from a T under the above modification rule, then* **GreedyHeuristic** *finds the original graph T.*

See [3] for details and for the description of the weighted version of the problem, where the edge weights represent degrees of similarity.

Experiments on protein data show that our algorithm is much faster than an exact algorithm (seconds instead of days runtime), while still giving optimal solutions for $> 75\%$ of the considered graphs. For incorrectly solved graphs, costs are, on average, within 102% of the optimum.

References

[1] M. Křivánek and J. Morávek. NP-hard problems in hierarchical-tree clustering. *Acta Informatica*, 23(3):311–323, June 1986.

[2] J. W. Moon. A note on approximating symmetric relations by equivalence classes. *SIAM Journal of Applied Mathematics*, 14(2):226–227, 1966.

[3] S. Rahmann, T. Wittkop, J. Baumbach, M. Martin, A. Truss, and S. Böcker. Exact and heuristic algorithms for weighted cluster editing. In: *Proceedings of 6th Conference on Computational Systems Bioinformatics*, pages 391–401, August 2007.

Analysis of Metabolite Tandem Mass Spectra

Sebastian Böcker,[1] Florian Rasche[2]

1 Introduction

Mass spectrometry is a high-throughput technology for the analysis of proteins and metabolites [2]. Since the manual interpretation of mass spectra is tedious, methods for automated analysis are highly sought. These methods may rely on databases. Because no databases are available for many applications and species, bioinformaticians have developed "de novo" interpretation methods capable of interpreting MS data with no need of any database.

We analyze metabolites using tandem mass spectra obtained from quadrupole time-of-flight mass spectrometers. In these devices the analyte is fragmented by collision induced dissociation (CID) [5]. As a first step of the de-novo analysis of metabolites, we identify the sum formula of the measured molecule.

2 Methods

We calculate the elemental decompositions for all fragments using the Round Robin-Algorithm [1]. We afterwards construct a colored graph using the decompositions as vertices and all possible fragmentation steps as edges. We assign the same color to vertices, if the corresponding decompositions belong to the same peak. Afterwards, we score decompositions and fragmentation steps using the following values: Peak intensity, mass deviation, hydrogen-to-carbon-ratio, and RDBE values [3]. Additionally, the score of well known neutral losses is increased. We avoid the use of strict filters to prevent sorting out a candidate too early. The scores may represent the likelihood that the corresponding decomposition or fragmentation step is correct.

From the graph constructed we calculate the most likely fragmentation tree. This tree is the Maximum Colorful Tree of the graph. Requiring the tree to be colorful avoids selecting two explanations for the same peak. Because this calculation is NP-hard, we use a fixed-parameter algorithm similar to Scott et al. [4] as well as heuristics to solve the problem.

3 Results

We tested our algorithms with metabolite spectra measured using an API QSTAR Pulsar Hybrid QTOF spectrometer by Applied Biosystems. These tests indicate that the proposed exact algorithm runs fast and produces good results. For all 45 compounds, six of them with a mass over 400 Da, the correct solution was among the top five suggestions, and for 39 compounds the first suggestion was correct. The greedy heuristic performed as good as the exact algorithm, whereas the top-down heuristic even improved the results. Tests on more data are necessary to validate this effect. Table 1 contains detailed results and Figure 1 shows a predicted fragmentation tree. All algorithms need about 1.5 minutes total running time to identify all 45 compounds.

4 Outlook

Zhang et al. [6] propose a tool that uses isotopic pattern of the fragments for identification. We already perform as well as Zhang et al. without using isotopic patterns. With a combination of both methods and data of more exact QTOF spectrometers we expect to be able to exactly identify the sum formula of all metabolites of mass up to 1000 Da exactly.

[1]Faculty of Mathematics and Informatics, University of Jena, Germany. Email: boecker@minet.uni-jena.de
[2]Faculty of Mathematics and Informatics, University of Jena, Germany. Email: florian.rasche@minet.uni-jena.de

References

[1] S. Böcker and Z. Lipták. A fast and simple algorithm for the money changing problem. *Algorithmica*, 48(4):413–432, 2007.

[2] Chhabil Dass. *Principles and Practice of Biological Mass Spectrometry*. John Wiley and Sons, 2001.

[3] T. Kind and O. Fiehn. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, 8(1):105, 2007.

[4] J. Scott, T. Ideker, R.M. Karp, and R. Sharan. Efficient algorithms for detecting signaling pathways in protein interaction networks. *J Comput Biol*, 13:133–144, 2006.

[5] J. M. Wells and S. A. McLuckey. Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol*, 402:148–185, 2005.

[6] J. Zhang, W. Gao, J. Cai, S. He, R. Zeng, and R. Chen. Predicting molecular formulas of fragment ions with isotope patterns in tandem mass spectra. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(3):217–230, 2005.

Mass range	# compounds	Exact			Greedy heuristic			Top-down heuristic		
		Top 1	Top 2	Top 5	Top 1	Top 2	Top 5	Top 1	Top 2	Top 5
100–200 Da	26	100%	100%	100%	100%	100%	100%	100%	100%	100%
200–300 Da	11	82%	82%	100%	82%	82%	100%	82%	90%	100%
300–400 Da	2	50%	100%	100%	50%	100%	100%	50%	100%	100%
400–500 Da	6	83%	83%	100%	83%	83%	100%	100%	100%	100%

Table 1: The identification rates of the exact FPT algorithm and the two heuristics.

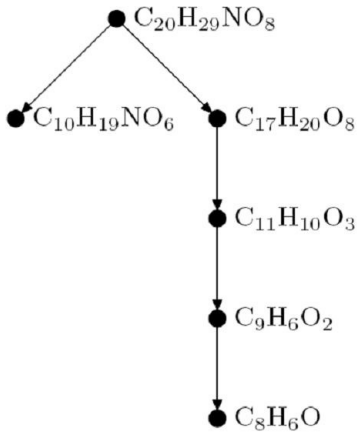


Figure 1: The predicted fragmentation tree of hexosyloxycinnamoyl choline.

L-Valine Production by Systematically Engineered *Escherichia coli*

Sang Yup Lee,[1,2,3*] Jin Hwan Park,[1,2]
Kwang Ho Lee,[1,2,4] Tae Yong Kim[1,2]

The L-valine producing strain of *Escherichia coli* was constructed by rational metabolic engineering and stepwise improvement based on transcriptome analysis and *in silico* gene knock-out simulation. Feedback inhibition of acetohydroxy acid synthase isoenzyme III by L-valine was removed by site-directed mutagenesis and the native promoter containing the transcriptional attenuator leader regions of the *ilvGMEDA* and *ilvBN* operon were replaced with the *tac* promoter. The *ilvA*, *leuA* and *panB* genes were deleted to make more precursors available for L-valine biosynthesis. This engineered Val strain harboring pKKilvBN, which overexpresses the *ilvBN* genes, produced 1.31 g/liter L-valine. Comparative transcriptome profiling combined with *in silico* gene knock-out simulation was used for the enhanced production of L-valine. The VAMF strain (Val $\triangle$*aceF* $\triangle$*mdh* $\triangle$*pfkA*) harboring pKBRilvBNCED and pTrc184ygaZHlrp was able to produce 7.55 g/liter L-valine from 20 g/liter glucose, resulting in a high yield of 0.378 g L-valine per g glucose. The approaches described here can be a good example of systematically engineering strains for the enhanced production of amino acids.

Acknowledgments. This work was supported by the Korean Systems Biology Project of the Ministry of Science and Technology (M10309020000-03B5002-00000). Further supports by the LG Chem Chair Professorship and KOSEF through the CUPS are appreciated.

[1]Metabolic and Biomolecular Engineering National Research Laboratory, Department of Chemical and Biomolecular Engineering (BK21 program), BioProcess Engineering Research Center, KAIST, Daejeon, Republic of Korea.
[2]Center for Systems and Synthetic Biotechnology, Institute for the Biocentury, KAIST, Daejeon, Republic of Korea.
[3]Department of BioSystems and Bioinformatics Research Center, KAIST, Daejeon, Republic of Korea.
[4]R & D Center for Bioproducts, CJ Corp., Seoul, Republic of Korea.
* Tel: 82-42-869-3930, Fax: 82-42-869-8800, Email: leesy@kaist.ac.kr

Deciphering the Evolution and Metabolism of *Mannheimia succiniciproducens* MBEL55E by Genome-Scale Analysis

Sang Yup Lee,[1,2,3] Tae Yong Kim[1,2]

This study presents an in-depth study on the organism behavior of *Mannheimia succiniciproducens*, the cell growth rate and succinic acid production rate, under varying rumen gas conditions. Constraints-based flux analysis of the genome-scale metabolic model of *M. succiniciproducens* was employed to estimate intracellular fluxes and the exchange fluxes across the cellular system associated with the metabolism of H_2 and CO_2. Results from fermentations performed previously and constraints-based flux analyses of *M. succiniciproducens* in this study revealed that there is a limit of CO_2 level in the medium for the increment in the cell growth rate. Furthermore, uptake rates of H_2 and CO_2 from the medium have a direct relationship with one another, significantly influencing the rates of cell growth and succinic acid production as a result.

Acknowledgments. This work was supported by the Genome-based Integrated Bioprocess Project of the Ministry of Science and Technology. Further supports by the LG Chem Chair Professorship, IBM SUR program, Microsoft, and by the KOSEF through the Center for Ultramicrochemical Process Systems are appreciated.

[1]Metabolic and Biomolecular Engineering National Research Laboratory, KAIST, 373-1 Guseong-dong Yuseong-gu, Daejeon 305-701, Republic of Korea.

[2]Department of Chemical and Biomolecular Engineering (BK21 program), KAIST, 373-1 Guseong-dong Yuseong-gu, Daejeon 305-701, Republic of Korea.

[3]Department of BioSystems, BioProcess Engineering Research Center and Bioinformatics Research Center, KAIST, 373-1 Guseong-dong Yuseong-gu, Daejeon 305-701, Republic of Korea. Tel: 82-42-869-3930, Fax: 82-42-869-8800, Email: `leesy@kaist.ac.kr`

A New Probabilisitic Approach for Simplified Partial Digest Problem

Duygu Taş,[1] Kemal Kılıç,[1] Osman Uğur Sezerman[1]

1 Introduction

Restriction site analysis and *hybridization* are used by molecular biologists to gain information from DNA molecules. In restriction site analysis, besides the length of the fragments, the cutting method is also provided as an input. The cutting method tcan be either *Double Digest* or *Partial Digest*. The computational complexity of the Partial Digest Problem leads to another method called *Simplified Partial Digest Problem*, which was introduced by Blazewicz et al. [1] as an alternative. Basically, two different reaction times are selected in the Simplified Partial Digest Problem. First reaction time is chosen in such a way that the DNA molecule is decomposed into fragments at one restriction site at most. On the other hand, the second reaction time is selected long enough to cut the DNA molecule at every restriction site. Abrams and Chen [2] propose an algorithm for Simplified Partial Digest Problem. The time complexity of the algorithm is $O(n \log n)$ and this algorithm finds a solution to the Simplified Partial Digest Problem with a probability which is higher than 0.5. The proposed algorithm is actually an extension of the algorithm proposed by Abrams and Chen. This new method is based on a probabilistic approach to solve the Simplified Partial Digest Problem which was not considered in previous studies. The DNA restriction map can be found by a forward-looking method which is used as a tie-breaker in our algorithm. The algorithm can also deal with the errors that usually occur during the measurement of the length of the fragments.

2 Method

The DNA fragments whose one endpoint is on the left or right side of the target DNA strand are called *primary fragments*. The fragments obtained from the long digestion process are called *base fragments*. In the proposed algorithm, a look ahead feature is incorporated which associates a probability to both of the possible assignments whenever there is a tie and the assignments are made based on these probabilities. In order to assign the probabilities, the proposed algorithm utilizes a look-ahead process. Let the look-ahead parameter be h. The algorithm exhaustively searches all possible combinations for the following h steps, i.e., 2^h possible combinations are created. Next, the number of feasible offspring is determined among the offspring in which the first assignment is made to the left side. The ratio of the number of feasible offspring to 2^{h-1} (total number of offspring in which the first assignment is made to the left side) is assigned as the probability of assigning the base fragment to the left hand side. Same thing is repeated for the right hand side and the corresponding probability is also obtained. Later these two probabilities are normalized so that the sum of them becomes one. Figure 1 depicts an example of resulting tree for the look ahead process with step size is equal to 3. After the probabilities are identified, the assignment is done randomly based on these probabilities.

One of the nice features of this random selection process is the fact that it allows different assignments each time the algorithm is repeated. Therefore one can repeat the algorithm as much as the time permits and eventually identifies a solution for any problem case. Also there might be multiple fragments with same size in the base fragment set, which does not constitute a problem in our algorithm.

Furthermore, our algorithm can handle the errors in the measurement of primary and base fragment lengths. The minimum and maximum differences of current minimum and the previous minimum primary fragments for both left and right sides are calculated taking into account error percentages resulting from length measurements. Also, the intervals of base fragments within the error value are calculated. If the interval of a base fragment intersects with the interval of the difference of one side (left or right), this base fragment can be assigned to the related side.

[1]Faculty of Engineering and Natural Sciences, Sabanci University, Turkey. Email: duygutas@su.sabaciuniv.edu, {kkilic, ugur}@sabanciuniv.edu

3 Experimental Analysis and Results

The algorithm either finds the correct DNA strand or stops if it can not find proper base fragment length from the list. Both ideal and noisy experiments were randomly generated by using the method explained in Blazewicz et al. [1]. Experiments were run on PC with Inter Core 2 CPU with 2.13 GHz and 0.97 GB of RAM.

The results of the experiments are presented in Table 1. The proposed algorithm, which is applied in MATLAB R2006b, can find the correct DNA molecule for 30 experiments out of 30 for error free cases. Also, for the noisy data our algorithm can find a solution for all the cases ($N = 10$, 16, and 20) depending on the number of forward-looking steps (h). Note that the probability of the algorithm defined in [2] finds a solution with a probability, that is to say with a probability at least 0.5. Furthermore, the computational time of the proposed algorithm outperforms the results that are presented in [1].

Overall, the presented algorithm is much faster than the approach presented in [1] and yields a result more often than the approach presented in [2]. The probabilistic nature of the proposed algorithm and the look-ahead feature seems to work very well for the simplified partial digest problem.

References

[1] Blazewicz J., Formanowicz P., Karprzak M., Jaroszewski M., Markiewicz W.T. Construction of DNA restriction maps based on a simplified experiment. *Bioinformatics*, 17(5):398–404, 2001.

[2] Abrams Z., Chen H. The simplified partial digest problem: Hardness and a probabilistic analysis. *Proc. of 4th Annual RECOMB Satellite Meeting on DNA Sequencing Technologies and Computation*, May 2004.

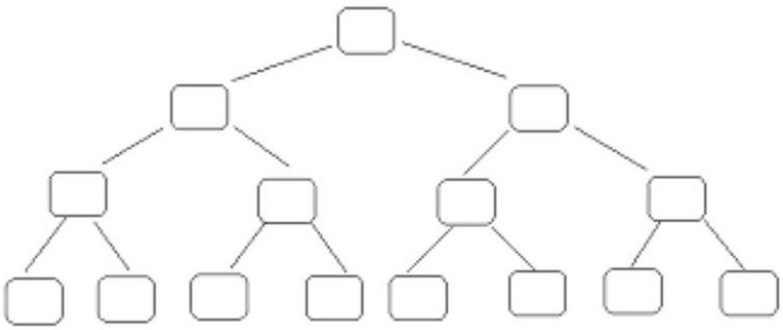


Figure 1: Forward steps in the SPDP.

Number of restriction sites (N)	Number of Experiments	Running time in seconds ($h = 2$)	Running time in seconds ($h = 3$)	Running time in seconds ($h = 4$)	Running time in seconds ($h = 5$)
10	4	0.0012	0.0012	0.0012	0.0011
16	3	0.0024	0.0024	0.0024	0.0028
20	5	0.0036	0.0036	0.0035	0.0036
40	2	0.0154	0.0155	0.0152	0.0150
60	2	0.0370	0.0376	0.0372	0.0375
100	4	0.1331	0.1354	0.1424	0.1589

Table 1: Some computational results for ideal data.

Constraint Programming Applied to Simplified Partial Digest Problem with Errors

Elvin Coban,[1] Kemal Kılıç,[1] Osman Uğur Sezerman[1]

1 Introduction

We developed a constraint programming (CP) model for the simplified partial digest problem (SPDP) in which an enzyme cuts a target DNA strand only at one site because of the time span allowed for the reaction. Besides handling perfect data cases, approaches to noisy data for measurement errors are done by modifying the CP model. After allowing the assignments of fragments, we have different objectives to evaluate the alternative solutions found by CP like minimizing the maximum error or minimizing the total error or minimizing the deviation from the estimated lengths. This research is the first one in which CP is used for SPDP. From the results we have obtained CP is a good declarative programming paradigm for this problem in terms of running times for data sets with errors and for large perfect data sets (when restriction site is over 60) when compared to existing results [1].

2 SPDP

In 1970 Hamilton Smith discovered HindII enzyme cleaves DNA molecule at every occurrence of the sequence GTGCAC or GTTAAC. After this, restriction maps became very important in molecular biology for achieving physical maps of the chromosomes [2, 3]. With these physical maps molecular biologists have a chance to map location of gene [4]. For mapping problems, Karp provides a good overview [5], and for restriction mapping Lander presents a good survey in his work [6].

Restriction mapping has a history of approximately four decades. There are different experimental approaches to this problem. One of them is partial digest approach, in which we are given all pairwise distances set and we try to form the order of DNA molecule such that we still satisfy the input data by the order we create. This problem is called also turnpike problem in computer science of forming geography of the highway given the input of every pair of exits on a highway from one town to another. The highway exits correspond to restriction sites of our partial digest problem. Only the used metrics differ in each problem, one is in miles and the other one is in nucleotides [3]. On the other hand, in the simplified partial digest problem, the enzyme cuts the DNA molecule either only at one restriction site at each experiment or at all sites. This is achieved by controlling the duration of the reaction. The length of the target DNA, multiset of distances of fragments when enzyme cuts at one restriction site and the multiset distances of fragments when the enzyme cuts at all restriction sites which are called bases are the input in the problem. The sequence of bases in proper order is what we are looking for.

Table 1 presents the variables used in the model and their ranges. In the model the length of the target DNA is called length. Note that $(0, \mathit{length})$ is also included as one of the pairings. For the noisy data case, p, the permitted error percentage which ranges generally from 2% to 7%, is used. The following CP model is used for the SPDP problem.

$$\frac{y_j + y_{k-j+1}}{1-p} \geq \frac{\mathit{length}}{1+p} \text{ and } \frac{\mathit{length}}{1-p} \geq \frac{y_j + y_{k-j+1}}{1+p}, \quad \forall j \in \{1, \ldots, m\} \tag{1}$$

$$\frac{x_j}{1-p} \geq \frac{y_{j+1}}{1+p} - \frac{y_j}{1-p} \text{ and } \frac{y_{j+1}}{1-p} - \frac{y_j}{1+p} \geq \frac{x_j}{1+p}, \quad \forall j \in \{1, \ldots, m-1\} \tag{2}$$

$$\frac{x+m}{1-p} \geq \frac{y_k}{1+p} - \frac{y_m}{1-p} \text{ and } \frac{y_k}{1-p} - \frac{y_m}{1+p} \geq \frac{x_m}{1+p} \tag{3}$$

$$y_1 = 0 \tag{4}$$

First constraint looks for if the assigned pairings' complementary elements sum up to the length of target DNA, in a sense whether they are feasible pairings or not. Second and third constraints look

[1]Faculty of Engineering and Natural Sciences, Sabanci University, Turkey. Email: elvinc@su.sabaciuniv.edu, {kkilic, ugur}@sabanciuniv.edu

for if the base fragment lengths can be obtained from the assigned pairings' complementary elements. Fourth constraint assigns the first pairing as 0 (therefore $y_k = length$). In the constraints the main idea is defining ranges considering the error percentage permitted. We do not include some of the right or left hand sides as they are already satisfied with the other constraint combined with the "and" constraint.

3 Results

In order to compare the performance of the CP in SPDP, we created data with different number of restriction sites based on the procedure detailed in Blazewicz et al. [1]. We created 5 sets of data for each restriction site number. Tests were run on PC with Inter Core 2CPU with 2.13 GHz and 0.97 GB of RAM and for handling CP we use ILOG OPL Studio 3.7.1. We consider permitted error ratio as 5%.

The comparison of the proposed CP approach with the existing algorithms is presented in Table 2. PDP and SPDP are the results presented by Blazewicz [1] and "SPDP with CP" is our results. When there are 100 restriction sites ($N = 100$), our model determine at least one solution in less than 8 seconds on the average. This is even shorter than the PDP's and SPDP's performance with 20 restriction sites ($N = 20$) with $p = 2\%$. Note that as p increases the search space of the algorithm also widens. For the previously proposed algorithms, the presented experiment with highest number of restriction sites is 20. Since we do not have a chance to compare our solutions when restriction site is 40, 60 and 100, we indicated the results with "*" for these runs in Table 2. On the other hand, in Table 2, "∞" refers to the case in which no solution is found in 5 minutes.

The proposed CP can determine at least one solution in a short time for all of the cases included in the experiments. This indicates the benefit of using CP for the problem compared to PDP and SPDP. Note that PDP does not yield a solution even for much smaller cases of $N = 12$ for $p = 1.5\%$ [1].

References

[1] Blazewicz J., Formanowicz P., Karprzak M., Jaroszewski M., Markiewicz W.T. Construction of DNA restriction maps based on a simplified experiment. *Bioinformatics*, 17(5):398–404, 2001.

[2] Setubal, J. C., Meidanis, J. *Introduction to Computational Molecular Biology*, Ch. 5, PWS Publishing Company, 1999.

[3] Jones., N. C., Pevzner, P. *An Introduction to Bioinformatics Algorithms*, Ch. 4, MIT Press, 2004.

[4] Shield, D.C., Butler, A., Mosurski, K. R., Walsh, M.T., Whitehead, A.S. Mapping genes within a YAC by computer-assisted interpretation of partial restriction digestions. *Nucleic Acids Res.*, 24(22):4495–4500, 1996.

[5] Karp, R. M. Mapping the genome: Some combinatorial problems arising in molecular biology. *Proc. 25th Annual ACM Symposium on Theory of Computing*, pages 278–285, 1993.

[6] Watermann, M. S. *Mathematical Methods for DNA Sequences.* pages 35–51, CRC Press, 1989.

m	Number of pairings* (number of base)	
k	Number of pairings' elements (2*m)	
Index:		
i	Defined for decision variable x	$1, \ldots, m$
j	Defined for decision variable y	$1, \ldots, k$
Decision Variables:		
x_i	Length of each base fragment	$1, \ldots, m$
y_j	Length of each pairings' element	$1, \ldots, k$

Table 1: Definition of variables and ranges.

$p =$	1.5%		2%		5%
Number of Restriction Site	PDP	SPDP	PDP	SPDP	SPDP with CP
10	0–25.5	≤ 0.01	0–152	≤ 0.01	0.04
16	∞	0–0.02	∞	0–0.10	0.102
20	∞	1.21–12.6	∞	1.10–83	0.1725
40	*	*	*	*	0.87
60	*	*	*	*	1.954
100	*	*	*	*	7.426

Table 2: Comparison of results of the runs in seconds of the noisy data with results of [1].

GACOT: A Genetic Algorithm for the Physical Mapping Problem with Noisy Data

Hsin-Nan Lin,[1] Wen-Lian Hsu[2]

1 Introduction

In DNA sequence analysis, the physical mapping problem is to determine the order of probes (or molecular markers) in a group of clones. The construction of physical maps is generally accomplished as follows. Long DNA sequences are separated into smaller fragments (called clones). A number of probes are tested for their presence or absence in the clones. Given the collection of probes each clone has been attached to, one tries to order the probes in such a way that probes belonging to the same clone are consecutive. The presence and absence of probes for a group of clones is represented by a 0-1 matrix. Theoretically, the 0-1 matrix for the physical mapping problem has the consecutive ones property, which means that there exists a column (probe) permutation such that the ones in each row (clone) of the resulting matrix are consecutive. The first linear time algorithm for consecutive ones test was proposed by Booth and Lueker . However, the algorithm would fail when the input matrix contains errors, which is quite common in wet-lab experiments.

The most common error types for the physical mapping problems are false positives, false negatives, non-unique probes, and chimeric clones. A false positive is an entry of 1 that should actually be 0. A false negative is an entry of 0 that should actually be 1. A non-unique probe is a probe sequence that occurs more than once in the DNA sequence. Two (or more) clones that incidentally stick together at their ends form a chimeric clone. Several related problems have been proved to be NP-hard. We develop a two-stage genetic algorithm, called GACOT, to tackle the physical mapping problems with synthetic errors, and compare it with a previous method proposed by Lu and Hsu (L&H) in 2003. The experiment results show that GACOT generates more accurate results than those of L&H's method.

2 Methods

The procedure of GACOT involves two major stages: arranging and connecting. In the first stage, we use a genetic algorithm to arrange the permutations of probes for each clone trying to find out the best ordering of probes which attach to that clone. At this stage, we do not decide the absolute position for each probe, but the relative position. For each clone, GACOT would find at least one probe permutation which makes the maximum consecutive ones for the sub-matrix related to that clone. In this stage, we also deal with the problem of false negatives by analyzing the sub-matrices of different probe permutations during the evolution process of the genetic algorithm. After all clones being processed in the first stage, we obtain the neighborhood information of each probe. The neighborhood information indicates which probes are neighbors to each others. Theoretically, each probe should have at most two neighboring probes without ambiguities. We remove those probes whose numbers of neighbors are greater than two before the second stage of GACOT.

In the second stage, we apply another genetic algorithm to connect the remaining probes according to the neighborhood information. Since the neighborhood information presents the relative positions of those probes, GACOT tries to find the best probe ordering which makes the longest connection among all remaining probes. The best probe ordering is the final answer for the physical mapping problem.

We conduct experiments on the synthetic data simulating these four error types. We use three fixed matrices of sizes 100x100, 200x200, and 400x400 that satisfy the consecutive ones property. These matrices are generated randomly under the constraint that the number of 1s in each row ranges from 5 to 15. The error rate of non-unique probes and chimeric clones are 2%, and the error rate of false positives and false negatives are generated at three different levels, 3%, 5%, and 10% respectively. Within each error level, the ratio of the number of false positives and that of false negatives is set to be 1 to 4. For each

[1]Institute of Information Science, Academia Sinica, Taiwan. Email: arith@iis.sinica.edu.tw
[2]Institute of Information Science, Academia Sinica, Taiwan. Email: hsu@iis.sinica.edu.tw

error combination generated, we repeat the experiment 100 times based on different random seeds. These experiment designs are proposed by L&H and we follow the same procedure to generate test dataset.

3 Results

We follow the evaluation method proposed by L&H and compare with their results, which is shown in Table 1. The results are evaluated by comparing the resultant column ordering from that of the original ordering using the measures defined below. For a column v, let d_1 be the number of columns ordered to the left of v but whose indices are greater than that of v and d_2, the number of columns ordered to the right of v whose indices are less than v. Let the displacement $d(v)$ of column v be the larger of d_1 and d_2. The displacement $d(v)$ gives an approximate measure of the distance of column v from its "correct" position. L&H defined the following three criteria for measuring the total deviation of the resultant ordering from the original one:

1. If the displacement of a column v is more than 4, we say v is a jump column. The jump percentage is the number of jump columns divided by the total number of columns.

2. The average displacement of a column ordering is the average of the displacement of all columns in the resultant order.

3. The average difference of the column ordering is the average of the difference in the column indices of adjacent columns in the resultant order.

According to the results, GACOT generate more accurate probe ordering than that of L&H. The number of jump columns is much less than L&H's results when the error rate is increasing. The average displacement and the average difference are also more robust when the error rate is increasing. It shows that GACOT is more reliable when there are more noises in the original data.

References

[1] Booth, K.S. and G.S. Lueker. Testing for Consecutive Ones Property, Interval Graphs, and Graph Planarity Using Pq-Tree Algorithms. *Journal of Computer and System Sciences*, 13(3):335–379, 1976.

[2] Lu, W.F. and W.L. Hsu. A test for the consecutive ones property on noisy data: Application to physical mapping and sequence assembly. *Journal of Computational Biology*, 10(5):709–735, 2003.

Dataset			100x100			200x200			400x400		
Error Case			3%	5%	10%	3%	5%	10%	3%	5%	10%
jump percentage	GACOT	=0%	99%	95%	94%	99%	93%	84%	95%	90%	76%
		≦5%	100%	100%	99%	100%	100%	100%	100%	100%	100%
		≦10%			100%						
		≦15%									
		≦20%									
	L&H	=0%	70%	50%	20%	38%	38%	26%	70%	60%	20%
		≦5%	100%	92%	90%	92%	92%	96%	100%	100%	98%
		≦10%		100%	96%	100%	96%	100%			98%
		≦15%			96%		98%				98%
		≦20%			100%		98%				100%
		≦25%					100%				
Average Displacement	GACOT		0.09	0.13	0.26		0.15	0.26	0.15	0.19	0.32
	L&H		0.02	0.06	0.37	0.05	0.44	0.77	0.92	0.82	1.01
Average Difference	GACOT		1.11	1.17	1.31	1.13	1.18	1.31	1.15	1.19	1.35
	L&H		1.39	1.60	1.97	1.56	1.68	2.07	1.46	1.56	1.99

Table 1: The performance comparisons between GACOT and L&H's results.

Probabilistic Arithmetic Automata and their Application to Pattern Matching Statistics

Tobias Marschall,[1,2] Sven Rahmann[2]

1 Introduction

Biological sequence analysis is often concerned with the search for structure in long strings like DNA, RNA or amino acid sequences. Frequently, "search for structure" means to look for patterns that occur very often. An important point in this process is to define sensibly a notion of "very often". One option is to consult the statistical significance of an event: Suppose we have found a certain pattern n times in a given sequence. What is the probability of observing n or more matches just by chance?

The topic of statistics of words on random texts has been studied extensively. An overview is provided in the book by Lothaire [4]; Chapter 6 ("Statistics on Words with Applications to Biological Sequences"), which is particularly interesting to us, is based on the overview article by Reinert et al. [7].

In most approaches developed until now, a generating function is derived for the sought quantity. Then, typically using symbolic Taylor expansion, the concrete values can be computed. Such a procedure is, for instance, described by Régnier [6], Nicodème et al. [5], and Lladser et al. [3].

We introduce the concept of *probabilistic arithmetic automata* (PAAs) and demonstrate how it paves the way for a dynamic programming approach to exact pattern matching statistics. The notion of PAAs can be seen as a generalization of *Markov additive chains* used by Kaltenbach et al. [2] for fragment statistics of cleavage reactions. A different dynamic programming approach was recently presented by Zhang et al. [10]. They use it to compute exact p-values for position weight matrices describing transcription factor binding sites.

2 Probabilistic Arithmetic Automata

Definition 3 (PAA) *A* probabilistic arithmetic automaton *is an 8-tuple* $(Q, T, q_0, N, n_0, E, \theta = (\theta_q)_{q\in Q}, \pi = (\pi_q)_{q\in Q})$*, where* Q *is a finite set of states,* $T : Q \times Q \to [0,1]$ *is a transition function, i.e.,* $(T(p,q))_{p,q\in Q}$ *is a stochastic matrix,* $q_0 \in Q$ *is called* start state*,* N *is a finite set called* value set*,* $n_0 \in N$ *is called* start value*,* E *is a finite set called* emission set*, each* $\theta_q : N \times E \to N$ *is an operation associated with the state* q*, and each* $\pi_q : E \to [0,1]$ *is a distribution associated with the state* q*.*

At first, the automaton is in its start state q_0, as for a classical deterministic finite automaton (DFA). In a DFA, the transitions are triggered by input symbols. In a PAA, however, the transitions are purely probabilistic; $T(p,q)$ gives the chance of going from state p to state q. While going from state to state, a PAA performs a chain of calculations on a set of values N. In the beginning, it starts with the value n_0. Whenever a state transition is made the entered state, say state q, generates an emission according to the distribution π_q. The current value and this emission are then subject to the operation θ_q, resulting in the next value. Let $\{Y_k\}_{k\in N_0}$ denote the automaton's state process and $\{V_k\}_{k\in N_0}$ the value process, that means the sequence of values resulting from the chain of performed calculations.

PAAs thus provide a formalization of computations on (conditional) probability distributions, in particular probabilities such as $P(Y_k = y, V_k = v)$ can be computed exactly. Here we explore the application of PAAs to pattern matching statistics.

3 Application to Pattern Matching Statistics

From a pattern, given in one form or another (e.g. a single string, a set of strings, a Prosite pattern,[3] a consensus string together with a distance measure and a distance threshold, an abelian pattern, a position

[1] This poster abstract is based on work Tobias Marschall did at Bielefeld University.

[2] Bioinformatics for High-Throughput Technologies at the Chair of Algorithm Engineering, Computer Science Department, TU Dortmund, 44221 Dortmund, Germany. Email: {tobias.marschall, sven.rahmann}@tu-dortmund.de

[3] Like used in the Prosite database; see Hulo et al. [1].

weight matrices in connection with a threshold, etc.), one can construct a deterministic finite automaton (DFA) recognizing that pattern. Together with a random text model, either i. i. d. or Markovian, we can construct the corresponding PAA. Then, using dynamic programming, we are able to exactly compute the joint distribution of states and occurrences of the given pattern.

3.1 Statistics of Protein Motifs from Prosite

Prosite is a database of biologically meaningful amino acid motifs; see Hulo et al. [1]. We use Prosite to assess the practicability of our method in the context of computational biology. The construction of PAAs succeeded for 96.8% of all prosite patterns (for 3.2%, the computation aborted due to memory limitations). For 94.9% of all patterns, the construction was finished within 2 seconds. The majority of resulting automata were of reasonable size; 91.9% of the patterns yielded automata with less than 10000 states and 79.5% with less than 500 states.

To give an impression of the runtimes to be expected in practice, consider an example pattern[4] from Prosite. It results in an automaton with 462 states. Computing the distribution of the occurrence count took 1 second (up to 50 occurrences, for a random text of length 1000).

3.2 Statistics of Transcription Factor Binding Sites

Transcription factor binding sites (TFBSs) are commonly represented by position weight matrices (PWMs). Therefore, it is an important task to compute the significance of occurrences of a PWM, for instance in a given promoter region. Using our framework, we have implemented two approaches to PWM statistics.

The first approach considers all strings whose PWM score is above a threshold as a match, enumerates them and builds the respective automaton. This approach is similar to that of Zhang et al. [10].

The second approach does not impose a hard threshold. Recently, Roider et al. [8] presented a procedure to predict a transcription factor's affinity to a sequence based on a physical model. Using their method, we estimate the probability that a TF binds to a particular sequence and incorporate these probabilities into our model.

In order to assess the practicability of both methods, we consulted the Jaspar database (see Sandelin et al. [9]). As a result, we found both methods to be applicable to the majority of motifs from Jaspar. Detailed results on this evaluation will be presented on the poster.

References

[1] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. Langendijk-Genevaux, M. Pagni, and C. Sigrist. The PROSITE database. *Nucleic Acids Research*, 34(S1):D227–D230, 2006.

[2] H.-M. Kaltenbach, S. Bocker, and S. Rahmann. Markov additive chains and applications to fragment statistics for peptide mass fingerprinting. In: *Systems Biology and Computational Proteomics*, LNCS 4532:29–41, 2006.

[3] M. Lladser, M. D. Betterton, and R. Knight. Multiple pattern matching: A Markov chain approach. *Journal of Mathematical Biology*, 56(1-2):51–92, January 2008.

[4] M. Lothaire. Applied Combinatorics on Words. In: *Encyclopedia of Mathematics and its Applications*, Cambridge University Press, 2005.

[5] P. Nicodème, B. Salvy, and P. Flajolet. Motif statistics. *Theoretical Computer Science*, 287:593–617, 2002.

[6] M. Régnier. A unifed approach to word occurrence probabilities. *Discrete Applied Mathematics*, 104:259–280, 2000.

[7] G. Reinert, S. Schbath, and M. S. Waterman. Probabilistic and statistical properties of words: An overview. *Journal of Computational Biology*, 7(1–2):1–46, 2000.

[8] H. Roider, A. Kanhere, T. Manke, and M. Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141, 2007.

[9] A. Sandelin, W. Alkema, P. G. Engström, W. W. Wasserman, and B. Lenhard. JASPAR: An open access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(1):Database Issue, 2004.

[10] J. Zhang, B. Jiang, M. Li, J. Tromp, X. Zhang, and M. Q. Zhang. Computing exact p-values for DNA motifs. *Bioinformatics*, 23(5):531–537, 2007.

[4] `C-x-H-R-[GAR]-x(7,8)-[GEKVI]-[NERAQ]-x(4,5)-C-x-[FY]-H`.

Metabolic Engineering of *Escherichia coli* for Production of Malic Acid

Tae Yong Kim,[1,2] Soo Yun Moon,[1,2] Soon Ho Hong,[3] Sang Yup Lee[1,2,4]

Malic acid is widely used as a specialty chemical with applications in polymers, foods and pharmaceuticals. Metabolic flux analysis was performed to find a strategy for enhanced malic acid production in *Escherichia coli*. The *in silico* simulation results suggested that the amplification of phosphoenolpyruvate (PEP) carboxylation flux allowed increased malic acid production. Since the PEP carboxylase of *E. coli* converts PEP to oxaloacetate without generating ATP, thus losing the high energy phosphate bond of PEP, the PEP carboxykinase which generates ATP during this conversion was chosen. However, the *E. coli* PEP carboxykinase catalyzes the reaction that converts oxaloacetate to PEP rather than the desirable opposite reaction. The *pta* mutant *E. coli* strain WGS-10 harboring the plasmid p104ManPck containing the *M. succiniciproducens pckA* gene was constructed. The final malic acid concentration of 9.25 g/L could be obtained after 12 h of aerobic cultivation.

Acknowledgments. This work was supported by the Genome-based Integrated Bioprocess Project of the Ministry of Science and Technology. Further supports by the LG Chem Chair Professorship, IBM SUR program, Microsoft, and by the KOSEF through the Center for Ultramicrochemical Process Systems are appreciated.

[1]Metabolic and Biomolecular Engineering National Research Laboratory, KAIST, 373-1 Guseong-dong Yuseong-gu, Daejeon 305-701, Republic of Korea.

[2]Department of Chemical and Biomolecular Engineering (BK21 program), KAIST, 373-1 Guseong-dong Yuseong-gu, Daejeon 305-701, Republic of Korea.

[3]School of Chemical Engineering & Bioengineering, University of Ulsan, 29 Mugeo-dong, Nam-gu, Ulsan 680-749, Republic of Korea.

[4]Department of BioSystems, BioProcess Engineering Research Center and Bioinformatics Research Center, KAIST, 373-1 Guseong-dong Yuseong-gu, Daejeon 305-701, Republic of Korea. Tel: 82-42-869-3930, Fax: 82-42-869-8800, Email: `leesy@kaist.ac.kr`

Metabolic Pathway Analysis and its Optimization for Producing Succinic Acid in *Mannheimia succiniciproducens* MBEL55E

Tae Yong Kim,[1,2] Hyung Rok Choi,[2] Sang Yup Lee[1,2,3]

Mannheimia succiniciproducens MBEL55E is a capnophilic gram-negative bacterium which efficiently produces succinic acid. In order to analyze metabolic pathways of *M. succiniciproducens*, we applied elementary-mode analysis to the biochemical network of *M. succiniciproducens*, previously developed by our group. In this biochemical network, reactions known to be inactive under anaerobic condition with glucose as a carbon source were removed from our research consideration. Because elementary-mode analysis is not applicable to the large-scale biochemical network, we only considered its central carbon metabolism. We then also analyzed the biochemical network of *Escherichia coli*, in the same way as above, in order to grasp the notable differences between these two organisms. In order to draw clear conclusions, we clustered the solutions of two microorganisms, and compared each other. Each of clusters showed characteristic yield of succinic acid and the number of solutions and clusters of *M. succiniciproducens* is greater than that of *E. coli*. The results manifested that *pckA* is the major factor of succinic acid promotion. This analysis can show the differences between networks of two organisms, and suggest efficient biochemical network design.

Acknowledgments. This work was supported by the Genome-based Integrated Bioprocess Project of the Ministry of Science and Technology. Further supports by the LG Chem Chair Professorship, IBM SUR program, Microsoft, and by the KOSEF through the Center for Ultramicrochemical Process Systems are appreciated.

[1]Metabolic and Biomolecular Engineering National Research Laboratory, KAIST, 373-1 Guseong-dong Yuseong-gu, Daejeon 305-701, Republic of Korea.

[2]Department of Chemical and Biomolecular Engineering (BK21 program), KAIST, 373-1 Guseong-dong Yuseong-gu, Daejeon 305-701, Republic of Korea.

[3]Department of BioSystems, BioProcess Engineering Research Center and Bioinformatics Research Center, KAIST, 373-1 Guseong-dong Yuseong-gu, Daejeon 305-701, Republic of Korea. Tel: 82-42-869-3930, Fax: 82-42-869-8800, Email: leesy@kaist.ac.kr

A Decision Support System for Cardiovascular Disease Using Bioinformatics Approach

Fang Rong Hsu,[1] Wei-Chung Shia[1]

1 Introduction

Cardiovascular disease (CVD) is one of top 10 causes of death in Taiwan. In this poster we have developed a medical decision support system to use in the healthcare and precaution of cardiovascular diseases, and it can help medical experts to do the prediction and estimate the progress and prognosis of CVD patient more accurate.

In this system, we integrate 34 cardiovascular diseases and its related gene expression data, SNP, protein-protein interaction and alternative splicing information into a web. Through analysis this data model in suitable algorithm, we will get the significant data and rules for the future research and tracking.

We wish this system can bring the new viewpoint to understand the cardiovascular disease, and provide better treatment to the patient.

2 Methods

Our goal is to establish a genetic database of cardiovascular disease. Therefore, we integrate these data and show on the web: The cardiovascular disease related gene and haplotype, alternative splicing and proteinprotein interaction (PPI). We also provide an interface to search and view these data together.

The source of cardiovascular disease related gene data is NCBI OMIM database [3]. First we collect all the OMIM disease data, and use the text-miming to generate the dataset of cardiovascular disease. Second, we analysis these dataset and get the list of the cardiovascular disease and its related gene. Finally, we parse these annotations and store in database, and filter the incorrect result.

The source of protein-protein interaction data is STRING [5]. The database STRING ('Search Tool for the Retrieval of Interacting Genes/Proteins') aims to collect, predict and unify most types of protein-protein associations, including direct and indirect associations. We use the cardiovascular disease related gene list to search the STRING database, and get the PPI network graph. Due to the cardiovascular disease is multicomplex disease, these graphs can help us to understand the all interactions of these related genes.

The source of alternative splicing data is AVATAR [2]. AVATAR is an add-value alternative splicing database. Alternative splicing is an important event of gene transcript, and it causes the polymorphism of the gene expression. We link this database and get the alternative splicing result of these cardiovascular disease related gene to help us to observe the form of specific gene.

The source of SNP data is HAPMAP [4]. HAPMAP provide plentiful SNP information, like the Linkage Disequilibrium (LD) Maps, tagSNPs and the classification data of race. We analysis these data and reserve the SNP data have high LD value that related the cardiovascular disease gene. These SNP data can help us to research the relation of specific SNP in specific race between the cardiovascular disease, and these data also help us to design the microarray experiment.

3 Result and Discussions

We have successfully built a system to reach our goal. Our system provides 34 cardiovascular diseases and its genetic data. Each disease all has the alternative splicing form graph, protein-protein interaction graph and related gene list and haplotype data.

In our system, the numbers of all CVD related genes are 480; the number of CVD related tagSNPs are 79621. Our website uses the PHP and GD to reach the search and visual interface, and we use mySQL to our database engine.

In the future, we will improve the function of system and add the microarray data display and analysis tool to provide researcher direct search the gene expression and SNP data. We also will proceed with the clinical experiment to verify our data, and get the precise result to improve the related research.

[1]Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan. Email: frhsu@fcu.edu.tw

References

[1] Fang Rong Hsu, Wei-Chung Shia, Jer Fong Chen, and Fang Ming Hsu. Single Nucleotide Polymorphism Mapping Using Multi-Layer Unique Markers. *Journal of Computers*, 18/3, 2007.

[2] Hsun-Chang Chang, Po-Shun Yu, Tze-Wei Huang, Fang-Rong Hsu, Yaw-Ling Lin. The Application of Alternative Splicing Graphs in Quantitative Analysis of Alternative Splicing Form from EST Database. *International Journal of Computer Applications in Technology*, 22(1):14–22, 2005.

[3] Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2007. `http://www.ncbi.nlm.nih.gov/omim`.

[4] The International HapMap Consortium. The International HapMap Project. *Nature*, 426:789–796, 2003.

[5] von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krger B, Snel B, Bork P. STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, 35(Database issue):D358–D362, 2007.

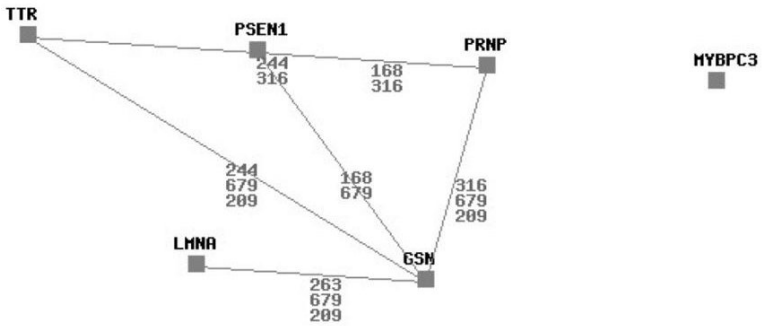


Figure 1: The PPI graph of the cardiac amyloidosis.

Disease	Gene Name
Aortic aneurysm	48
Arrhythmogenic right ventricular cardiomyopathy	22
Arterial thromboembolic disease	13
Ascending aortic disease	28
Atherosclerotic vascular disease	48
Brugada syndrome	6
Cardiac amyloidosis	9
Cardiomyopathy familial restrictive	26
Carney complex	21
Carnitine palmitoyltransferase II deficiency, late-onset form	2
Cerebral amyloid angiopathy	26
Congenital sick sinus syndrome	6
Coronary disease	212
Digeorge syndrome	79
Dilated cardiomyopathy	122
Familial hypercholesterolemia	76
Familial hypertrophic cardiomyopathy	78
Infantile dilated cardiomyopathy	18
Insulin resistance-related hypertension	21
Jervell and Lange-Nielsen syndrome	3
Myocardial infarction	146
Naxos disease	4
Orthostatic hypotension	31
Polymorphic ventricular tachycardia	21
Venous thrombosis	40
Ventricular tachycardia	46
Watson syndrome	156
Williams syndrome	514

Table 1: The list of cardiovascular disease we provided and its number of related gene.

Metabolic Engineering of *Escherichia coli* for L-Threonine Production based on Systems Biology

Sang Yup Lee,[1,2,3] Kwang Ho Lee,[1,2,4]
Jin Hwan Park,[1,2] Tae Yong Kim[1,2]

Amino acid producers have traditionally been developed by repeated random mutagenesis owing to the difficulty in rationally engineering the complex and highly regulated metabolic network. By combined genome engineering, transcriptome analysis, and genome-scale metabolic flux analysis, we report the development of the first genetically-defined L-threonine (Thr) overproducing *Escherichia coli* strain. All known feedback inhibitions, transcriptional attenuation regulations, and those pathways that negatively affect Thr production were removed by genome engineering. Several target genes were identified by transcriptome profiling combined with flux response analysis, and were engineered accordingly. The final engineered *E. coli* strain was able to produce 82.4 g/l Thr by fed-batch culture. The strategy of systems metabolic engineering reported here can be employed for developing genetically-defined organisms for the efficient production of bioproducts.

Acknowledgments. This work was supported by the Korean Systems Biology Project of the Ministry of Science and Technology (M10309020000-03B5002-00000). Further supports by the LG Chem Chair Professorship and KOSEF through the CUPS are appreciated

[1]Metabolic and Biomolecular Engineering National Research Laboratory, Department of Chemical and Biomolecular Engineering (BK21 program), BioProcess Engineering Research Center, KAIST, Daejeon, Republic of Korea.
[2]Center for Systems and Synthetic Biotechnology, Institute for the Biocentury, KAIST, Daejeon, Republic of Korea.
[3]Department of BioSystems and Bioinformatics Research Center, KAIST, Daejeon, Republic of Korea. Tel: 82-42-869-3930, Fax: 82-42-869-8800, Email: leesy@kaist.ac.kr
[4]R & D Center for Bioproducts, CJ Corp., Seoul, Republic of Korea.

Application of Genome-Scale Metabolic Model of *Vibrio vulnificus* CMCP6 for *In Silico* Drug Targeting

Tae Yong Kim,[1] Hyun Uk Kim,[1] Joon Haeng Rhee,[2] Sang Yup Lee[1,3]

1 Introduction

Vibrio vulnificus is a halophilic and highly human-pathogenic bacterium, showing very high mortality rate when infected [1]. In order to facilitate the drug development process for this, we undertook in silico analysis to identify specific drug targets in the genome-scale metabolism of *V. vulnificus*. With a newly sequenced and annotated genome of *V. vulnificus*, we first reconstructed its genome-scale metabolic network consisting of 946 reactions and 766 metabolites (Table 1). Subsequently, we employed constraints-based flux analysis [2], an optimization-based simulation technique, to validate the model in comparison with experimental data, and identify essential genes comprising the metabolic network. Essential genes herein refer to genes responsible for the specific enzymatic reactions whose deletions result in the failure of biomass formation.

In order to identify primary drug targets, we applied constraints-based flux analysis to the genome-scale model of *V. vulnificus* with maximization of biomass as an objective function under random media. Here, the random media indicate a set of media covering all possible combinations of carbon and nitrogen sources so as to account for various nutrients available for the pathogens inside the human body. In this study, the random media consist of 22 carbon sources and 41 nitrogen sources, and the simulation was performed for each combination. Uptake of sulfate, phosphate and oxygen was allowed in all cases. As a result, 228 enzymatic reactions were identified as primary drug targets. This study demonstrates that drug targeting using *in silico* approaches facilitates not only the systems-level analysis of the bacterial metabolism, but also a rational design of experiments applicable to biomedical science.

Acknowledgments. This work was supported by the Korean Systems Biology Project of the Ministry of Science and Technology (M10309020000-03B5002-00000). Further supports by the LG Chem Chair Professorship, Microsoft, and IBM SUR program are appreciated.

2 Software and Files

The genome-scale metabolic model was validated by comparing simulation results with those from the literature as detailed below, and accordingly further refined so as to use it as a source of *in silico* drug targeting. The model was constructed using MetaFluxNet [3], and it was converted into GAMS (GAMS Development Corp., Washington DC, USA) for subsequent drug targeting simulations.

Features	Number
Genome feature:	
Genome size (base pairs, bp)	5,126,798
No. of open reading frames (ORFs)	4,796
In silico metabolic model:	
No. of reactions (redundant) included in model	946
No. of biochemical reactions	810
No. of transport reactions	136
No. of metabolites	766
No. of ORFs assigned in metabolic network	669

Table 1: The genome and the *in silico* genome-scale stoichiometric model of *V. vulnificus* CMCP6.

References

[1] Gulig, P.A., Bourdage, K.L. and Starks, A.M. 2005. Molecular Pathogenesis of Vibrio vulnificus. *Journal of Microbiology*, 43:118–131.

[2] Price, N.D., Reed, J.L. and Palsson, B.O. 2004. Genome-scale models of microbial cells: Evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2:886–897.

[3] Lee, D.Y., H. Yun, S. Park, and S.Y. Lee. 2003. MetaFluxNet: The management of metabolic reaction information and quantitative metabolic flux analysis. *Bioinformatics*, 19:2144–2146.

[1]Department of Chemical and Biomolecular Engineering (BK21 Program), Metabolic and Biomolecular Engineering National Research Laboratory, Center for Systems and Synthetic Biotechnology, Institute for the BioCentury, KAIST, Daejeon 305-701, Republic of Korea.

[2]Clinical Vaccine R & D center, Chonnam National University Medical School, Kwangju 501-746, Republic of Korea.

[3]Department of Bio and Brain Engineering, BioProcess Engineering Research Center and Bioinformatics Research Center, KAIST, Daejeon 305-701, Republic of Korea.

Pathogen Discovery by Combination of Computational Substraction and Pyrosequencing Technology

Roel G.W. Verhaak,[*,1,2] Laura MacConaill,[*,1,2] Carsten Russ,[1] Jen Chen,[1,3] Brian Desany,[4] Danny A Milner Jr,[3] Matthew Meyerson[1,2]

1 Introduction

Many diseases, including some cancers, inflammatory diseases and autoimmune disorders, have a suspected infectious etiology [1, 2, 3]. In the past, a number of microbiological and molecular methodologies have been applied to detect microbes or viruses associated with human disease. Successful examples include amplification of conserved sequences using polymerase chain reaction (PCR) and application of a DNA microarray containing highly conserved viral sequences [4, 5]. A limitation of these approaches is the dependence on known and conserved sequence. We have developed an unbiased method for pathogen discovery, computational subtraction, which is based on the assumption that diseased tissue should contain both host genomic DNA (or RNA) as well as nucleic acid from the causative infectious agent [6]. A computer-based comparison aligns sequence from diseased tissue to database human genome sequence(s) and subsequently subtracts matching sequence. Theoretically, only the disease-causing sequence should remain following the subtraction process. Here, we show the feasibility of this approach and apply it on seven different disease tissues from variable pathobiological backgrounds.

2 Materials and Methods

The following samples of autoimmune disease were collected

- Rheumatoid arthritis (n = 5, processed on quarter plates)
- Multiple sclerosis (n = 1)
- Giant cell myocarditis (n = 1)

A number of samples of tumor tissue were selected, including

- Chronic lymphocytic leukemia (n = 1)
- Hodgkins lymphoma (n = 1)
- Squamous cell carcinoma (n = 1)
- Lung adenocarcinoma (cell line) (n = 1)

A breast cancer cell line positive for Epstein-Barr virus was included as positive control (n = 4).

Massively parallel pyrosequencing was performed on these samples (Margulies, *Nature*, 2005). Computational subtraction was applied in which all sequences were aligned against various reference genomes, filtering all sequences matching with scores above threshold. Presence of remaining sequences in the original sample was confirmed through PCR.

[1]The Broad Institute of Harvard and MIT, Cambridge, MA, USA.
[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA.
[3]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts, USA.
[4]454 Life Sciences Corp. Branford, CT 06405, USA.
* Contributed equally. Email: verhaak@broad.mit.edu

3 Results

11,25 sequencing experiments resulted in sequence with a length of 242 Mb. Aligning against human genome build 36.1 resulted in removal of 99.997% of the sequences. Interestingly, only 2% of all EBV sequences present at start were removed. Aligning the residue against various other recently published human genomes resulted in a further removal of sequences. A number of remaining sequences aligned against various pathogen genomes. However, since PCR did not confirm the presence of these sequences original sample it is likely that these pathogens are the result of contamination sources in the experimental pipeline.

4 Discussion

Here, we present an approach that has previously been shown to work but which is now also practically feasible through the emergence of new high throughput sequences technologies. The loss of 2% EBV sequences in the positive control versus a 99.997% reduction of human sequences shows the effectiveness of the methodology. Although no pathogens detected in our experiments were confirmed in PCR experiments, this again shows the underlying strategy is robust and practical.

The quest for new pathogens is challenged by the spatiotemporal requirements of sample collection; a pathogen could be present at time of onset but not at time of detection. In the case of autoimmune disease a pathogen could be present systemically but induce a tissue specific response, leading to collection of the wrong tissue. However, with the diminishing costs and further advancement of high throughput sequencing technology, the identification of disease causing agents has become ever more likely.

References

[1] Relman DA. The search for unrecognized pathogens. *Science*, 284(5418):1308–1310, 1999.

[2] Chang Y. Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science*, 266(5192):1865-1869, 1994.

[3] Walboomers JM, Jacobs MV, Manos MM, Bosch FX, Kummer JA, Shah KV, Snijders PJ, Peto J, Meijer CJ, Munoz N. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *Journal of Pathology*, 189(1):12–19, 1999.

[4] Nikkari S, Gotoff R, Bourbeau PP, Brown RE, Kamal NR, Relman DA. Identification of Cardiobacterium hominis by broad-range bacterial polymerase chain reaction analysis in a case of culture-negative endocarditis. *Archives of Internal Medicine*, 162(4):477–479, 2002.

[5] Wang D, Urisman A, Liu YT, Springer M, Ksiazek TG, Erdman DD, Mardis ER, Hickenbotham M, Magrini V, Eldred J, Latreille JP, Wilson RK, Ganem D, DeRisi JL. Viral discovery and sequence recovery using DNA microarrays. *PLoS Biology*, 1(2):E2, 2003.

[6] Weber G, Shendure J, Tanenbaum DM, Church GM, Meyerson M. Identification of foreign gene sequences by transcript filtering against the human genome. *Nature Genetics*, 30(2):141–142, 2002.

[7] Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

EPoS: A Modular Framework for Phylogenetic Analysis

Thasso Griebel,[1] Malte Brinkmeyer,[2] Sebastian Böcker[3]

There exist several software packages for phylogenetics analysis, but they usually have shortcomings in either usability or support for computational methods. In particular supertree construction requires modularity as well as direct feedback for computation and analysis. EPoS is a modular software framework for phylogenetic analysis, visualization, and data management. It provides a plugin based system that integrates a storage facility, a rich user interface, and the ability to easily incorporate new method and functions. Its plugin fundament allows EPoS to extend in every direction. New algorithms, tools and visualizations can easily be added and the already existing modules can be extended. EPoS ships with persistent data management, visualizations, and a set of well known phylogenetic algorithms. Implemented methods cover distance based tree construction, consensus trees and various graph based supertree methods. EPoS supports rich tree visualizations embedded in a user-friendly interface. Several tree layouts are supported and the rendering system can be customized for, say, different edge and node styles. Executables and source code are available under the LGPL license at `http://www.bio.informatik.uni-jena.de/epos`.

1 Introduction

EPoS is a modular software framework that supports data management, computational methods and visualizations for phylogenetic analysis. Existing phylogenetic software tools attend to a wide range of applications. They support computational methods, visualizations and database integration and cover the area of computational phylogenetics comprehensively. Problems mostly occur in usability aspects of almost all of the available tools. Algorithmic packages are often command line based and enforce a good understanding of the software environment, while visualization tools suffer from poor graphical user interfaces and data integration abilities without providing sophisticated support for integrated computational methods. Furthermore, most programs rest upon their own, unique file formats, which makes data exchange between the programs difficult.

EPoS fills this gap by combining a powerful graphical user interface with a plugin system that allows simple integration of new algorithms, visualizations and data structures. It offers a simple way to incorporate new modules into the framework, without limiting the modularization to certain areas. In fact, the system itself is build from a set of core modules, which allows extensions in all directions. Limitations concern only the graphical user interface (GUI) and interaction model. The consistent EPoS GUI is used to manage and store data as well as to start available computational methods. Thus the workflow is disconnected from the data or applied methods and the module system prevents plugins from manipulating this structure to assure that the common work ow stays untouched. This, however, does not apply to visual extension for data analysis. Visualizations are part of the core system and can be extended in any direction.

2 Visualizations

EPoS already contains a comprehensive tree view that offers different layouts, colorization, annotation, and export functions, but the framework offers the ability to integrate views on all kinds of data. New visualization modules can be integrated to handle, say, phylogenetic networks. The integrated tree view module focuses on interactive tree analysis and provides functionality to display even large trees up to a few thousand leaves, without loosing the ability to smoothly interact with the view. Interactions are not restricted to one view. The underlying data model also allows communication between different views.

[1]Email: `thasso@minet.uni-jena.de`
[2]Email: `malte@minet.uni-jena.de`
[3]Chair of Bioinformatics, Faculty of Mathematics and Computer Science, University of Jena, Germany. Email: `boecker@minet.uni-jena.de`

For example, EPoS contains a method to compare the structure of two trees. This employes the ability to manipulate one view from another and allows side by side analysis of two trees. The model used to allow such interactions is based on the data management facility intergrated into EPoS.

3 Data Management

To simplify data handling, EPoS creates a persistent workspace that contains all used data sets. Data within the workspace are persistently stored in a transparent and extendable backend module. For example, a tree is stored in a data object and a user opens a view on that tree. A data object for the view is created and linked to the tree data. It is used to store the visual configuration (colors, annotations, layouts), while changes to the tree structure are delegated and stored directly in the tree. The core module that is responsible for the data management uses an embedded database as default storage location. This assures both simplicity and extensibility. The user does not have to manually manage the database. Actually, he does not even have to know about it. It is automatically started and used by the framework. In contrast, the mechanism that maps EPoS data objects to the relational database table is transparent, such that one can use the same storage process on a local or remote database server without changing the data objects.

The data objects provide another feature that encourages extensibility. All persistently stored data objects carry their private data and provide the additional ability to store further properties of any kind. For example, when annotating data, Web Services can be used to correlate different data sets and obtain additional information, without modifying the data objects implementation. The data are stored as key-value pairs within the object. This is also used by some of the computational methods that need supplementary information. An implementation of Ranked Tree [3] is integrated into EPoS. This method needs additional information about the input trees, in this case, information about divergence dates. They are added as an additional property directly to the trees. This simplifies both data management and execution of the algorithm.

4 Methods

Currently, Ranked Tree, as well as all other methods, are integrated into a pipeline system. It allows combinations of methods that are executed sequentially, where the data flow is handled automatically by the system. EPoS provides pipelines for different computational methods. It supports distance based tree reconstruction methods including Neighbor Joining and Agglomerative Clustering, consensus construction, such as Adams- and N-Consensus, and several supertree methods that construct trees from overlapping leave sets. EPoS directly supports Aho's Build [1], MinCut [5], modified MinCut [4], Ranked Tree [3], and Ancestral Build [2] as graph-based supertree algorithms. No external software packages have to be installed to use one of these algorithms.

The execution environment is another extendable part within the framework. EPoS uses the local machine as the default location to execute pipelines, but it is not limited to the local environment. In combination with the persistence mechanism, data can easily be moved to other machines or compute grids, and the execution environment can be shifted as well. This will allow the system to move computationally intensive processes to a remote machine or a cluster.

References

[1] Alfred V. Aho, Yehoshua Sagiv, Thomas G. Szymanski, and Jeffrey D. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM J. Comput.*, 10(3):405–421, 1981.

[2] Vincent Berry and Charles Semple. Fast computation of supertrees for compatible phylogenies with nested taxa. *Syst. Biol.*, 55(2):270–288, 2006.

[3] Semple C. Bryant, D. and M. Steel. Supertree methods for ancestral divergence dates and other applications. Pages 129–150, Computational Biology Series, Kluwer, 2004.

[4] Roderic D. M. Page. Modified mincut supertrees. In: *Proc. Workshop on Algorithms in Bioinformatics* (WABI 2002), LNCS 2452:537–552, 2002.

[5] Charles Semple and Mike Steel. A supertree method for rooted trees. *Discrete Appl. Math.*, 105(1-3):147–158, 2000.

A New Triad Based Approach to Sequence Comparison of Various Types of Collagen

S. Avinash Kumar,[1] S. Sundar Raman,[2]
R. Parthasarathi,[2] V. Subramanian[2]

1 Introduction

Collagen is an important protein due to its presence in the human body in large quantity as well as its biomedical and commercial applications. 28 different types of collagen have been identified so far. Numerous studies have been carried out on various aspects of collagen including structure, function, diseases, development of biomaterials and its commercial applications. But much still remains to be learned about this protein. In particular further studies on the functional and structural differences between the various types of collagen are necessary in order to understand the differences between them and their relationships.

2 The Triad Concept

It is desirable to study differences between the types of collagen at the level of their amino acid sequences itself. Conventional sequence analysis of collagen provides information about importance of various triplets in the stabilization of collagen. However such a sequence analysis does not include information about the interaction between the three strands that make up the collagen triple helix. Our new sequence comparison strategy takes this into account by including the amino acids from all the three polypeptide strands that make up the collagen triple helix in the form of a "TRIAD". Based on the information generated from this approach, proposed models for various types of collagen have been computationally derived. The usefulness of the triads in the analysis of various types of collagen will be presented.

References

[1] Bhattacharjee A., Bansal M. Collagen structure: The Madras triple helix and the current scenario. *IUBMB Life*, 3:161–172, 2005.

[2] GN Ramachandran, M Bansal, RS Bhatnagar. A hypothesis on the role of hydroxyproline in stabilizing collagen structure. *Biochim Biophys Acta*, 322(1):166–171, 1973.

[3] Ramachandran GN. Structure of Collagen. *Nature*, 177:710–711, 1956.

[1]School of Biotechnology, Chemical and Biomedical Engineering, Vellore Institute of Technology University, Vellore 632 014, Tamil Nadu, India. Email: avinash.rsg@gmail.com
[2]Chemical Laboratory, Central Leather Research Institute, Adyar, Chennai 600 020, Tamil Nadu, India.

Structural Classification Using Mining of Frequent Patterns in Concave Protein Surfaces

Sumeet Dua,[1,2] Shirin A. Lakhani,[1] Hilary W. Thompson[2]

1 Introduction

Protein structural classification, specifically *in silico* functional annotation of proteins is an overriding problem in the field of bioinformatics. Classifying proteins based on sequential and structural features using the conventional methods is arduous and inaccurate, partially due to the weak representation of the protein subunits that provide the discriminatory behavior. Interest in classifying proteins using protein surface information has grown in recent years. Research interest in protein surface regions, specifically the concave surfaces has grown because these areas provide specialized regions of biological activity. Well-formed concave surface regions can therefore be examined to identify similarity relationships in proteins.

2 Approach

In this work, we propose a new association rule based technique using concave residues and residue parameters of proteins to find the frequent spatial arrangement of residue, which is unique to a particular family of proteins. The first step in this technique is to discover association rules for all classes of proteins [1] that satisfy minimum support and minimum confidence constraints for class-level rule discovery and appraisal. The second step in this technique is to use Classification Based Association (CBA) rule mining [2] to discover frequent patterns that are present on the concave protein surfaces and that will aid in the discovery of a small set of rules satisfying minimum support and minimum confidence. The outline of our approach is presented in Figure 1.

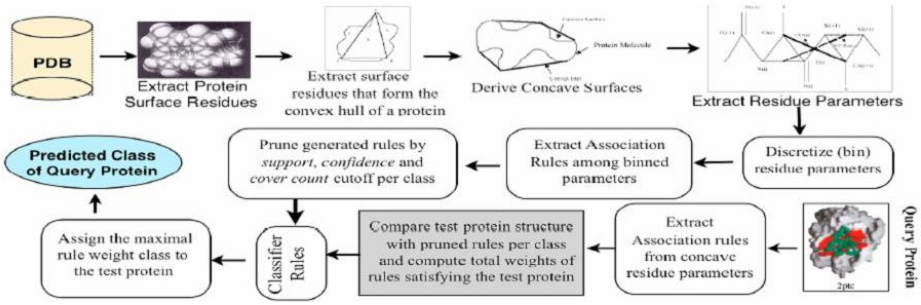


Figure 1: Outline of the proposed methodology.

We observe that association rules based framework yield better results than other classification techniques. We also discover and use the item-sets (attribute aggregates of protein surface residues) or residue parameters that are frequent for a class. Rules that satisfy minimum thresholds are extracted and employed for classification purposes. A query protein is subjected to the rule extraction method and compared with unique rules generated during the training phase. The protein is classified into a structural class whose rules most satisfy the protein features with enhanced degrees of specificity and sensitivity.

3 Results

In order to compare and analyze the classification approach of association rule mining on the concave protein surface, we conducted two sets of experiments. The first experiment, on fold level classification,

[1]Data Mining Research Laboratory, Dept. of Computer Science, Louisiana Tech University, Ruston, LA 71272, USA. Email: sdua@coes.latech.edu
[2]Department of Ophthalmology, Louisiana State University Health Sciences Center, New Orleans, LA 70112 USA.

was conducted by comparing 15 folds. The second experiment, on family level classification, was conducted using three protein families. We evaluated the accuracy performance of our proposed framework by conducting an experiment on 600 proteins randomly selected from 15 folds (in the SCOP hierarchy [3]). These proteins, with less than 40% sequence homology, were selected from the FSSP database. In this experiment, the training dataset contained 540 proteins (36 from each of the 15 folds), and the test dataset contained 60 proteins (4 from each fold). The concave surfaces for each of the 540 proteins were extracted, and patterns common to the proteins in each fold were derived in the form of classassociation rules particular to each fold. The classification was then conducted using the rules that satisfied the minimum threshold of 2% support and 60% confidence. Figure 2 shows the fold-wise and overall accuracy results on the 60 test protein dataset.

Effect of minimum support threshold (MST): The value of the minimum support for association rule discovery is critical. If the MST is set too high, we may not find those rules that involve the class in which we are interested. On the other hand, if the MST is set too low, there may be a combinatorial explosion because the majority class may have too many rules, most of which are over-fitted with many conditions and cover very few data cases. Figures 4 and 5 illustrate the true positive rates obtained for each class at various ranges of MST for a few selected classes.

The second experiment was performed on a dataset of 269 protein structures belonging to three protein structural families (Figure 3). The SCOP families that we classified include the Nuclear receptor ligand binding domain (NRLB) family from the all alpha proteins class, the Prokaryotic serine protease (PSP) family from the all beta proteins class, and the Eukaryotic serine protease (ESP) family, also from the all beta proteins class. Three datasets for the pair-wise comparison and classification of the above families were constructed. We randomly selected 228 proteins (60% of the original dataset) to represent the training set and the remaining 151 proteins (40%) formed the test dataset. Five random data sets were generated using the holdout method. The average classification accuracy rate using our approach was computed from a series of accuracy rates obtained from such iterations on the dataset (Figure 3).

Fold (Fold ID as in SCOP)	Class (as in SCOP hierarchy)	Ave Accuracy (in %)
46688	(1) All-α	100.00
47472	(2) All-α	75.00
48370	(3) All-α	75.00
48725	(4) All-β	100.00
50198	(5) All-β	41.67
51350	(6) α / β	91.67
51734	(7) α / β	100.00
51904	(8) α / β	83.33
52171	(9) α / β	33.33
52539	(10) α / β	100.00
52832	(11) α / β	83.33
53066	(12) α / β	83.33
53473	(13) α / β	75.00
54235	(14) $\alpha + \beta$	100.00
54861	(15) $\alpha + \beta$	91.67
Overall		82.22

Figure 2: Fold-wise and overall accuracy results on the 60 test protein dataset.

Classification tasks	Ave Accuracy (in %)
NRLB vs. PSP	86.36
ESP vs. PSP	92.73
RP vs. SP	92.55

Figure 3: Accuracy for pair-wise classification of selected SCOP families.

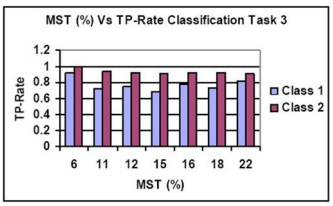


Figure 4: Effect on TPR with changing values of MST for classes 1 & 2 in Figure 2.

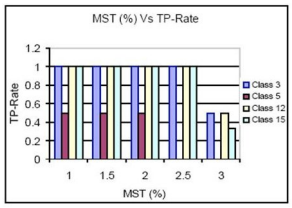


Figure 5: Effect on TPR with changing values of MST for classes 3, 5, 12 & 15.

4 Conclusions

Recent studies in computational protein structural classification have focused on developing better machine learning methods that could boost classification accuracy using less emphasis on the novelty of the structural descriptors themselves. In this work, we uniquely employ the residue position slopes and intercepts of the concave residue as the basis of a novel association rule discovery approach. The classifier that we use finds frequent patterns that maximize the intra-class similarity and reduce the inter-class dependencies of protein features. The experimental results have demonstrated superior sensitivity and specificity and open several interesting directions for further scientific pursuit.

References

[1] Lakhani S. Protein structural classification using mining of frequent patterns in concave protein surfaces, *MS Thesis*, Dept. of Computer Science (Advisor: Dua S.), Louisiana Tech University, Ruston, LA, May, 2007.

[2] Liu, B., Hsu, W., and Ma, Y. Integrating Classification and Association Rule Mining. *Knowledge Discovery in Databases*, 1998.

[3] Murzin, A., G., Brenner, S., E., Hubbard, T., and Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540. 1995.

HLA Class I Peptides: Exploiting Positional Information for Identification and Classification

Ankit Rakha,[1] Mitra Basu,[2,3] Rao Kosaraju[4]

1 Introduction

An essential primary step to stimulate cytotoxic T-cell response is via presentation of endogenous antigenic peptides, typically of viral origin, on a cell surface by Human Leukocyte Antigen (HLA) Class I molecules.[5] Thus, the ability to identify peptides that can bind to HLA molecules is of practical immunological importance. The polymorphism in HLA is concentrated around nucleotides encoding peptides adjoining the HLA peptide-binding groove. Each distinct HLA allele encodes a slightly different peptide-binding domain. A large almost distinct spectrum of peptides bind each particular HLA molecule. It is theorized that HLA binding peptides share certain binding patterns. Peptides that bind to HLA class I molecules are 7–12 amino acids long. Because of large HLA allelic variations, a systematic wet laboratory approach to prepare a T-cell epitope catalogue, even for a single protein antigen, requires a very large number of experiments. Since T-cell epitopes are a subset of HLA-binding peptides, identification and prediction of peptides that bind to HLA could be used effectively towards preselection of potential T-cell epitopes.

Individual residues of amino acid for a 9-mer peptide are named P1, ..., P9 beginning at the amino terminal end (P1) and finishing at the carboxyl end (P9). A certain number of amino acids out of these nine, bind to HLA playing the roles of primary and secondary anchors. A subset of the rest bind to T-cell to form the HLA-peptide-T cell complex if this peptide is a member of the set of T-cell epitopes.

A variety of methods such as structural motif-based approach; machine learning approach that includes support vector machine, artificial neural network and hidden Markov model; combination approaches that exploit biological properties along with algorithmic approach has been proposed for prediction of peptide binding to HLA [2, 4, 7, 8]. Binding motifs that exclusively rely on anchor position(s) have shown to be unreliable for classification/prediction purpose, indicating the participation of other positions in HLA-peptide binding process. Bowness [1] with some wet lab experiments describes that HLA-B27 most likely uses positions P2 (primary anchor), P3 and P9 (secondary anchors) for peptide binding. Recent studies report possible role of specific amino acid at each of these nine positions [5]. Others have relied on position specific scores (individual or pair-wise) to study HLA-peptide binding [4].

In this paper we study the importance of individual and combination of anchor positions, irrespective of the amino acid that occurs at that position, of a peptide in the context of HLA-peptide binding. Note that, the set of peptides that we experiment with contains epitopes as a subset. Our experiments produce some interesting byproducts that may be relevant to HLA-peptide complex and T-cell binding scenario. Our hypothesis is that positional importance can be extracted from data using some form of global learning technique. Our initial findings largely agree with previously reported wet lab experiments. We report some new observations regarding anchor positions. The findings reported here apply to the general classes of HLA-A and HLA-B.

2 Materials and Method

We consider two classes of 9-mer peptides: (i) 1255 peptides that bind to HLA-A and (ii) 323 peptides that bind to HLA-B (collected from `http://www.jenner.ac.uk`). Each amino acid is represented by a 5×1 property vector [6] producing a 45-dimensional vector for each peptide. So, the data is a set of points in 45-dimensional property space. We split the dataset randomly for training (70%) and testing (30%) and create 3 such sets. We use Adaboost algorithm [3, 9] with a variation of weighted nearest-neighbor classifier for the weak-learner part. Classification results on three test sets are shown in Table 1.

Next, we remove one, two and three[6] position-specific amino acids (to be called p-amino acid)) and study the respective classification results (see Table 2). Note that the classes now contain either 8-mer,

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. Email: `ankit21@cs.jhu.edu`
[2]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. Email: `basu@cs.jhu.edu`
[3]Department of Computer Science, United States Naval Academy, Annapolis, MD, USA. Email: `basu@usna.edu`
[4]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. Email: `kosaraju@cs.jhu.edu`
[5]HLA class I molecules are encoded at HLA-A, HLA-B and HLA-C loci positions.
[6]All combinations are considered.

7-mer or 6-mer peptides. Removal of p-amino acid presumably collapses boundary between HLA-A and HLA-B classes as property space dimension decreases. For validation of the fact that the proposed method performs poorly (i.e., it fails to learn) when presented with data where class boundary does not exist, we devise two sets (by picking randomly): Set-1 (50% of HLA-A and 50% of HLA-B) and Set-2 (the remaining HLA-A and HLA-B). We run the proposed method on 9 such 2-set datasets and name this experiment as Random Data Classification Experiment.

3 Results and Discussion

The averages of classification % for 9 datasets in the Random Data Classification Experiment are 63.8% and 53.0% for Set-1 and Set-2 respectively, which is low compared to the correct classification rate of (aproximately) 88%. Table 1 shows correct classification % for HLA-A and HLA-B for three 9-mer peptide test datasets. In Table 2, we only report results for combinations containing position 2 since other combinations do not show such significant drop. The first row shows overall drop in classification % with position 2 removed. The next two rows show the range of % drop in overall correct classification % when one or two other positions are removed along with position 2. We make the following observations.

1. All positions are important to an extent since removal of any single position reduces the % of correct classification. In other words, each position plays a role either in HLA-peptide binding or in binding between HLA-peptide complex and T-cell.

2. Position 2 is an anchor position and its removal alone reduces significantly the distinction between HLA-A and HLA-B. The largest drop (39.56%) in classification rate occurs when positions 2, 3 and 9 are removed.

3. Lowest drop in classification rate occurs for combinations that exclude positions 2 and 9 and contain positions 4 and 8. Perhaps this indicates that the positions 4 and 8 are possible actors in the HLApeptide complex and T-cell binding drama.

Our next step will use a modified version of Adaboost to provide a more holistic as well as quantitative measure of HLA-peptide interaction.

References

[1] Bowness, P. 2002. HLA B27 in health and disease: A double edged sword? *Rheumatology*, 41, pp. 857–868.

[2] Doytchinova, I., Flower, D. R. 2006. Class I T-cell epitope prediction: Improvements using a combination of proteomic cleavage, TAP affinity, and MHC binding. *Molecular Immunology*, 43, pp. 2037–2044.

[3] Freund, Y. and Schapire, R. E. 1996. A decision-theoretic generalization of on-Line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, pp. 119–139.

[4] Peters B., Tong, W. et al. 2003. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics*, 19, pp. 1765–1772.

[5] Pissurlenkar, R. R. S., Malde, A. K. et al. 2007. Encoding type and position in peptide QSAR: Application to peptides binding to class I MHC molecule HLA-A*0201. *QSAR Combinatorial Science*, 26, pp. 189–203.

[6] Venkatarajan, M. S., Braun, W. 2001. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. *Journal of Mol. Model*, 7, pp. 445–453.

[7] Yanover, C., Hertz, T. 2005. Predicting protein-peptide binding affinity by learning peptide-peptide distance functions. In: *RECOMB 2005*, pp. 456–471.

[8] Yu, K., Petrovsky, N. et al. 2002. Methods for prediction of peptide binding to MHC molecules: A comparative study. *Molecular Medicine*, 8, pp. 137–148.

[9] Zhu, J., Rosset, S. et al. 2006. Multi-class Adaboost. `http://www-stat.stanford.edu/ hastie/Papers/samme.pdf`

Test Set	HLA-A	HLA-B
1	84.35	92.78
2	88.06	86.60
3	84.62	90.72

Table 1: Classification rate (9-mer peptides).

Position Removed	Drop in Correct Classification %
Position 2	24.00
Position 2 & one other	19.4 – 38.63
Position 2 & two others	18.20 – 39.56

Table 2: Drop in classification rate with positions removed.

Protein Design by Sampling an Undirected Graphical Model of Residue Constraints

John Thomas,[1] Naren Ramakrishnan,[2] Chris Bailey-Kellogg[3]

Protein engineering seeks to produce amino acid sequences with desired characteristics, such as specified structure [1] or function [4]. This is a difficult problem due to interactions among residues; choosing an amino acid type at one position may constrain the possibilities at others, in order for the resulting protein to have proper structure and activity. To account for the dependence of some residues and take advantage of the independence of others, we have developed a new approach to protein design based on undirected probabilistic graphical models (Fig. 1). Our approach first constructs a graphical model that encodes residue constraints, and then uses the model generatively to produce new sequences optimized to meet the constraints. We focus here on constraints due to residue coupling, common pairs of amino acid types at particular pairs of positions, also known as correlated mutations or co-evolving residues. Recently, Ranganathan and colleagues showed that accounting for residue coupling, in addition to conservation, was to some extent both necessary and sufficient for viability of new WW domains [5, 6].

We have previously developed an approach for learning an undirected graphical model encapsulating conservation and coupling constraints in a protein family [7]. Our model provides a formal probabilistic semantics for reasoning about amino acid choices, defining a probability distribution function measuring how well a new sequence satisfies coupling constraints observed in the extant sequences of a family. Thus in order to design high-quality novel sequences, we can optimize for their likelihood under the model. Furthermore, our model explicates dependence and independence relationships between residue positions, so that we may reason about the impact of an amino acid choice at one position on those at others.

While sampling from an undirected model is difficult in general, we have developed two complementary algorithms that effectively sample the constrained sequence space. *Constrained shuffling* generates a fixed number of high-likelihood sequences that are reflective of the amino acid composition of a given family. A set of shuffled sequences is iteratively improved so as to increase their mean likelihood under the model. *Component sampling* explores the high-likelihood regions of the space and yields a user-specified number of sequences. Sequences are generated by sampling the cliques in a graphical model according to their likelihood, while maintaining neighborhood consistency. In contrast to the approach used by Ranganathan and colleagues, which simply seeks to reproduce the aggregate degree of coupling without regard to the quality of the individual sequences, our methods utilize a graphical model to generate sequences that meet the observed constraints, thereby improving the chances the designed sequences will be folded and functional. Theoretical results show that both of our methods properly sample the underlying sequence distribution.

We have applied our sampling algorithms in a study of WW domains, small proteins that assist in protein-protein interactions by binding to proline rich targets. We first showed that likelihood under our graphical model, trained on 42 wild-type WW domains, is predictive of foldedness for the new sequences designed by Ranganathan and colleagues, achieving a classification power of 0.8. We then generated novel putative WW domains optimizing the predicted likelihood. Both methods generated sequences with likelihoods near those of the wild-type WW domains, while being relatively novel and diverse (Fig. 2). The designed sequences serve as hypotheses for further biological study.

Our learning and sampling methods are applicable to a wide variety of protein families that may be targets for protein design. While multiple sequence alignments provide fundamental information on sequence constraints, our models may also incorporate additional structural or functional information. By including functional subclass information [7], we can design proteins with specific functional properties. By incorporating energetic constraints on side-chain interactions [2] we can design proteins with favorable predicted free energy.

[1]Department of Computer Science, Dartmouth College, Hanover, NH, USA. Email: jthomas@cs.dartmouth.edu
[2]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA. Email: naren@cs.vt.edu
[3]Department of Computer Science, Dartmouth College, Hanover, NH, USA. Email: cbk@cs.dartmouth.edu

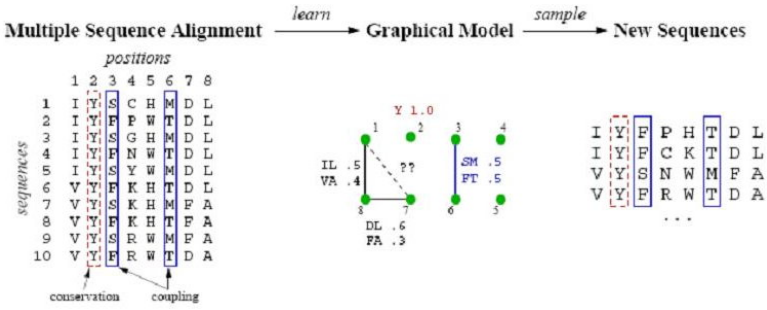


Figure 1: Given a multiple sequence alignment for a protein family (left), conservation and coupling constraints are inferred and summarized into a graphical model (middle) which captures conditional independence relationships through its edges. Through its clique potentials (not shown here), the model captures probability distributions for subsets of residues. Sampling from the model (right) then yields new sequences that obey the underlying constraints.

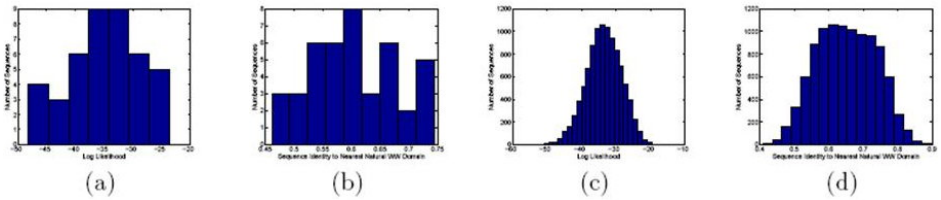


Figure 2: The log likelihood distribution (a, c), and sequence identity to the nearest natural WW domains (b, d), for the 42 and 10000 sequences generated by constrained shuffling and component sampling, respectively. The average log likelihood scores for the designed sequences are -34.69 with a standard deviation of 6.46 (constrained shuffling) and -33.48 with a standard deviation of 5.02 (component sampling). The wild-type WW domain sequences have an average log likelihood score under the model of -32.65 with standard deviation 4.93.

References

[1] B.I. Dahiyat and S.L. Mayo. De novo protein design: Fully automated sequence selection. *Science*, 278(5335):82–87, Oct 1997.

[2] H. Kamisetty, E.P. Xing, and C.J. Langmead. Free energy estimates of all-atom protein structures using generalized belief propagation. In: *Proc. RECOMB2007*, pages 366–380, Apr 2007.

[3] S.W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, Oct 1999.

[4] L.L. Looger, M.A. Dwyer, J.J. Smith, and H.W. Hellinga. Computational design of receptor and sensor proteins with novel functions. *Nature*, 423(6936):185–190, May 2003.

[5] W.P. Russ, D.M. Lowery, P. Mishra, M.B. Yaffee, and R. Ranganathan. Natural-like function in artificial WW domains. *Nature*, 437(7058):579–583, Sep 2005.

[6] M. Socolich, S.W. Lockless, W.P. Russ, H. Lee, K.H. Gardner, and R. Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–518, Sep 2005.

[7] J. Thomas, N. Ramakrishnan, and C. Bailey-Kellogg. Graphical models of residue coupling in protein families. *IEEE Trans. Comp. Biol. and Bioinf.*, 2007. In press. Preprint available at `http://www.cs.dartmouth.edu/~cbk/papers/tcbb07.pdf`.

A Novel Algorithm for Tag SNP Selection based on Pair-Wise Linkage Disequilibrium

Wei Wang,[1] Youling Guo,[1] Yuexian Zou,[1] Tianrui Wu[1]

The search for the association between complex diseases and single nucleotide polymorphisms (SNPs) or haplotypes has recently received great attention. For those studies, it is essential to use a small subset of informative SNPs, i.e., tag SNPs, accurately representing the rest of the SNPs. We describe an efficient algorithm for tagSNP selection based on pair-wise LD measure r2. The algorithms were implemented in a computer program named PLEXT (Partition-Ligation Exhaustive Search for Tagging) with MATLAB language. We first break down large marker sets into separate partitions, where more exhaustive searches can replace the LDSelect algorithm for tagSNP selection. Our algorithm leads to smaller tagSNP sets being generated. Its performance was assessed using HapMap data.

References

[1] G. C. L. Johnson, L. Esposito, B. J. Barratt, et al. Haplotype tagging for the identification of common disease genes. *Nature Genetics*, vol. 29, pp. 233-237, Oct 2001.

[2] S. B. Gabriel, S. F. Schaffner, H. Nguyen, et al. The structure of haplotype blocks in the human genome. *Science*, vol. 296, pp. 2225-2229, Jun 2002.

[3] D. A. Hinds, L. L. Stuve, G. B. Nilsen, et al. Whole-genome patterns of common DNA variation in three human populations. *Science*, vol. 307, pp. 1072-1079, Feb 2005.

[4] X. Y. Ke, S. Hunt, W. Tapper, R. Lawrence, et al. The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Human Molecular Genetics*, vol. 13, pp. 577-588, Mar 2004.

[5] C. S. Carlson, M. A. Eberle, M. J. Rieder, et al. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics*, vol. 74, pp. 106-120, Jan 2004.

[6] K. Zhang, M. H. Deng, T. Chen, et al. A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci USA*, vol. 99, pp. 7335-7339, May 2002.

[7] D. O. Stram, C. A. Haiman, J. N. Hirschhorn, et al. Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the multiethnic cohort study. *Human Heredity*, vol. 55, pp. 27-36, 2003.

[8] D. B. Goldstein, K. R. Ahmadi, M. E. Weale, and N. W. Wood. Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends in Genetics*, vol. 19, pp. 615-622, Nov 2003.

[9] Z. Lin and R. B. Altman. Finding haplotype tagging SNPs by use of principal components analysis. *American Journal of Human Genetics*, vol. 75, pp. 850-861, Nov 2004.

[10] B. V. Halldorsson, V. Bafna, R. Lippert, et al. Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. *Genome Research*, vol. 14, pp. 1633-1640, Aug 2004.

[11] A. Rinaldo, S. A. Bacanu, B. Devlin, et al. Characterization of multilocus linkage disequilibrium. *Genetic Epidemiology*, vol. 28, pp. 193-206, Apr 2005.

[12] D. Altshuler, L. D. Brooks, A. Chakravarti, et al. A haplotype map of the human genome. *Nature*, vol. 437, pp. 1299-1320, Oct 2005.

[13] D. E. Reich, M. Cargill, S. Bolk, et al. Linkage disequilibrium in the human genome. *Nature*, vol. 411, pp. 199-204, May 2001.

[14] E. Dawson, G. R. Abecasis, S. Bumpstead, et al. A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, vol. 418, pp. 544-548, Aug 2002.

[15] R. Sedgewick and P. Flajolet. *An Introduction to the Analysis of Algorithms*. Beijing: China Machine Press, 2006.

[16] B. Devlin and N. Risch. A comparision of lingkage disequilibrium measures for fine-scale mapping. *Genomics*, vol. 29, pp. 311-322, Sep 1995.

[17] X. Y. Ke, M. M. Miretti, et al. A comparison of tagging methods and their tagging space. *Human Molecular Genetics*, vol. 14, pp. 2757-2767, Sep 2005.

[1]Key Laboratory of Integrated Microsystems, Shenzhen Graduate School of Peking University. Email: wangw05373@szcie.pku.edu.cn

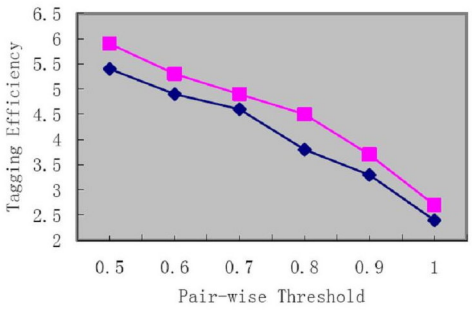


Figure 1: Tagging efficiency in the ENm010 region of CHB samples using the two tagging methods. Tagging efficiency was defined as the number of genotyped markers divided by the number of tagging SNPs. Lines with pink squares denote the PLEXT algorithm, lines with blue diamonds denote the LDSelect algorithm.

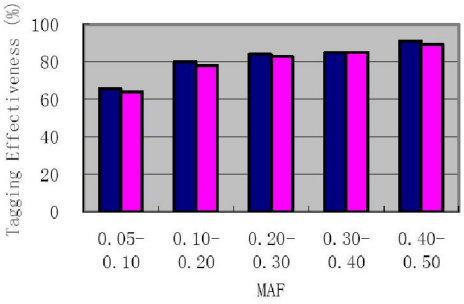


Figure 2: Tagging effectiveness in the ENm010 region of CHB samples using the two tagging methods at a fixed pair-wise threshold. Tagging effectiveness was defined as the percentage of hidden SNPs which had LD correlations with tagging SNPs over a threshold (r2 =0.5). Pink bars denote the PLEXT algorithm, blue bars denote the LDSelect algorithm.

	CEU	YRI	JPT+CHB
No.of SNPs			
r2 $\geq$ 0.5			
No.of partitions (No.of tag $>$ 2)	11782	24736	10249
Singletons	5342	15036	4316
No.of tagSNPs (ldSelect)	14376	27799	12456
No.of tagSNPs (PLEXT)	12750	26360	11102
r2 $\geq$ 0.8			
No.of partitions (No.of tag $>$ 2)	23435	41094	20163
Singletons	11341	22450	8958
No.of tagSNPs (ldSelect)	24300	41705	21033
No.of tagSNPs (PLEXT)	22950	39630	19943

Table 1: Summary of chromosome 7: Size of separate partitions and number of SNPs and Tag-SNPs in each partition.

CHB Samples	ENm010	ENm013	ENm014	ENr112	Enr 113
No.of SNPs	602	1376	1613	1096	1035
r2 $\geq$ 0.5					
No.of partitions	53	89	97	62	54
singletons	54	52	48	27	43
No.of tagSNPs (ldSelect)	111	139	143	91	93
No.of tagSNPs (PLEXT)	102	128	131	82	85
Reduction rate	15.8%	12.6%	12.3%	14.1%	16.0%
r2 $\geq$ 0.8					
No.of partitions	111	133	142	111	71
singletons	50	100	112	60	74
No.of tagSNPs (ldSelect)	158	223	245	160	144
No.of tagSNPs (PLEXT)	135	205	223	142	132
Reduction rate	20.6%	14.6%	16.5%	18.0%	20.0%

JPT Samples	ENm010	ENm013	ENm014	ENr112	Enr 113
No.of SNPs	629	1384	1615	1119	1034
r2 $\geq$ 0.5					
No.of partitions	75	103	128	104	69
singletons	73	97	104	59	62
No.of tagSNPs (ldSelect)	152	203 236 161 133			
No.of tagSNPs (PLEXT)	135	183	213	146	122
Reduction rate	21.5%	18.9%	15.2%	14.7%	15.5%
r2 $\geq$ 0.8					
No.of partitions	55	74	96	68	50
singletons	46	38	40	25	27
No.of tagSNPs (ldSelect)	103	116	134	90	79
No.of tagSNPs (PLEXT)	92	104	126	79	68
Reduction rate	19.3%	15.4%	8.5%	16.9%	21.1%

Table 2: Summary of Tag-SNPs identified by the ldSelect algorithm, the PLEXT algorithm in the five ENCODE regions of CHB and JPT samples.

Exact P-value calculation for clusters of TFBSs. Application in Computational Annotation of Regulatory Sites

Valentina Boeva,[1] Julien Clément,[2] Mireille Régnier,[3]
Mikhail A. Roytberg,[4,5] Vsevolod J. Makeev[6,7]

1 Introduction

In eukaryotic genomes a regulatory region is often bound by multiple transcription factors. Even one transcription factor can have multiple binding sites within one regulatory module [3]. The fact that motifs of binding sites form such dense clusters is widely used by a number of motif finding tools. Among them one can mention ClusterDraw, Stubb, MotifScan, SeSiMCMC, etc. All these predictive tools exploit the fact that a cluster of binding sites implies regulation. However, occurrence of multiple motifs of many factors together with their possible overlapping and fuzziness may complicate the assessing of statistical significance of an observed motif configuration. The compound Poisson distribution formula for the p-value could provide a good approximation, but not in the case of multiple highly overlapping motifs. As to overlapping motifs, especially within a heterotypic cluster, there one needs more precise method for statistical significance evaluation. In [1] we proposed an exact algorithm for p-value calculation for the general case of heterotypic clusters of motifs, which was implemented in AhoPro software, `http://favorov.imb.ac.ru/ahokocc`. Here we demonstrate how the above p-value approach may be used in annotation of DNA sequences according to transcription regulation. In particular, we made more precise maps of TF binding sites for transcription factors regulating the early development of *D. melanogaster* and showed the change of PWM threshold values in case of strong and 'shadow' motifs. The p-value method is applicable to find out transcription factors interacting with each other, and to discover regulatory regions controlled by several transcription factors.

This study has been supported by INTAS grant 05-1000008-8028, RFBR grant 07-04-01584 and by Russian Academy of Sciences MCB project.

2 Methods and Results

We present the method for assessing transcription regulation that consists in a two step procedure. First, given a DNA region and a position weight matrix corresponding to a motif for binding site of a transcription factor under consideration, we calculate the number of motif occurrences for each value of threshold. Obviously, the smaller is the threshold the higher is the number of motif occurrences. At the second step, we compare those numbers with the numbers of occurrences that could be observed by chance in a random sequence. E.g., for the lowest value of the threshold, we would find the same number of occurrences in any sequence of the same length. The best way of such a comparison with number of occurrences in a random sequence is the p-value calculation. The exact p-value computation, which allows for possible motif overlaps, was realized using the AhoPro tool [1]. For a given number of motif occurrences in a sequence S of length N and a threshold T, p-value is a probability to find at least the same number of motifs scoring higher than threshold T in a random sequence of length N with the same letter distribution as in S. Then, given the all p-values for all threshold values and looking at the whole p-value 'curves' (Fig. 1) one can judge about possible regulation. The advantage of this approach comparing to the approach presented in [2] is that here we can take into account simultaneous (possibly overlapping) occurrences of binding sites for several different factors. Besides, low score sites even of the same position weight matrix can frequently overlap, which influences the p-value. Thus, here we can consider, at the

[1]Laboratory of Computer Science, École Polytechnique, Palaiseau, France. Email: `valeyo@yandex.ru`
[2]GREYC, CNRS UMR 6072, Laboratory of Computer Science, Caen, France. Email: `Julien.Clement@info.unicaen.fr`
[3]INRIA Rocquencourt, Le Chesnay, France. Email: `Mireille.Regnier@inria.fr`
[4]Puschino State University, Puschino, Moscow Region, Russia. Email: `mroytberg@mail.ru`
[5]Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Puschino, Moscow Region, Russia.
[6]Institute of Genetics and Selection of Industrial Microorganisms, GosNIIGenetika, Moscow, Russia. Email: `makeev@genetika.ru`
[7]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia.

same time, strong and 'shadow' putative binding sites, and any intermediates. In Figure 1 we present such 'curves' for two sets: (A) a set of *cis*-regulatory regions with experimentally confirmed regulation by *D. melanogaster* early development transcription factor *bicoid*; and (B) a set of *cis*-regulatory regions in the same gene regulatory network but with the absence of confirmed regulation. The ordinate axis shows the negative logarithm of the p-values corresponding to different threshold values. As one can see, the p-values are significantly smaller nearly for all threshold values in the case of regions really regulated by *bicoid*, which means that we observe clusters of binding sites of various strengths in these DNA sequences. The minimal p-values are shown with circles. In Figure 1 we demonstrate that it is possible to distinguish between regulated and nonregulated regions and to find thresholds corresponding to clusters of strong and 'shadow' sites. The similar procedure can be used to assess regulation by cooperatively or competitively binding factors. Idea of the approach for heterotypic clusters was presented in [1].

3 Figures

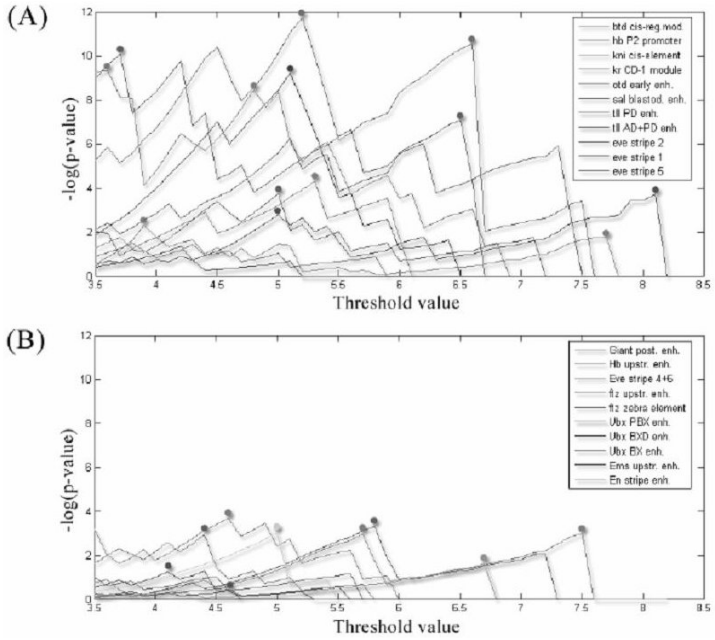


Figure 1: P-value curves for different threshold values for DNA regions regulated (A) and non-regulated (B) by *bicoid*. For each of 21 *cis*-regulatory regions (11 regulated by *bicoid*, 10 non-regulated) we searched for the number of motif occurrences scoring higher than given threshold values; then, calculated the p-value to find such number of motif occurrences in a random sequence.

References

[1] Boeva, V., Clement, J., Régnier, M., Roytberg, M.A., Makeev, V.J. 2007. Exact p-value calculation for heterotypic clusters of regulatory motifs and its application in computational annotation of cis-regulatory modules. *Algorithms for Mol. Biol.*, pp. 2:13.

[2] Lifanov, A.P., Makeev, V.J., Nazina, A.G., Papatsenko, D.A. 2003. Homotypic regulatory clusters in Drosophila. *Genome Research*, 13(4):579–588.

[3] Papatsenko, D.A., Makeev, V.J., Lifanov, A.P., Régnier, M., Nazina, A.G., Desplan, C. 2002. Extraction of functional binding sites from unique regulatory regions: The Drosophila early developmental enhancers. *Genome Research*, 12(3):470–481.

Residue Interaction Networks for Analyzing Resistance Mutations in HCV Protein Structures

Mario Albrecht,[1] Christoph Welsch,[1,2] Francisco S. Domingues,[1] Gabriele Mayr,[1] Andreas Schlicker,[1] Stefan Zeuzem,[2] Thomas Lengauer[1]

1 Introduction

A variety of computational approaches has used residue interaction networks for identifying critical amino acids in protein structures [1, 3]. The corresponding two-dimensional graphs commonly consist of edges that connect residue nodes and represent non-covalent interactions of amino acids. Here, we analyze the residue interaction network of the hepatitis C virus (HCV) protease NS3-4A. More than 170 million people worldwide are chronically infected with HCV and are at risk of developing liver cirrhosis and hepatocellular carcinoma [5]. In particular, we focus on the new protease inhibitor telaprevir, which showed a substantial antiviral effect in patients infected with HCV genotype 1 during a phase 1b clinical trial [4, 6]. This trial found residue mutations that confer varying degrees of drug resistance. Specific mutations at the protease positions V36 and T54 were associated with low to medium levels of drug resistance during viral breakthrough, resulting in an intermediate reduction of viral replication fitness.

2 Results and Discussion

Using available HCV crystal structures of the NS3-4A protease, we study the binding mode of different ligands including telaprevir [7]. Since V36 and T54 are located in the protein interior and far away from the ligand-binding pocket, we construct a network of non-covalent interactions of protease residues, including interacting ligands. The residue interactions are formed by hydrogen bonds and van der Waals forces based on the protease crystal structure. To suggest residues of structural or functional importance, we compute network topology parameters using our Cytoscape plugin NetworkAnalyzer [2]. We describe the potential impact of V36 and T54 mutants on side chain and backbone conformation and on residue interactions. We also discover possible molecular mechanisms for the observed mutational effects on antiviral drug activity and viral fitness. T54 mutants may affect the viral replication efficacy to a larger degree than V36 mutants when interfering with the catalytic triad and the ligand-binding site of the protease. V36 and T54 mutations may result in impaired protease residue interactions with the cyclopropyl group of telaprevir, leading to viral breakthrough variants.

References

[1] Amitai, G. et al. 2004. Network analysis of protein structures identifies functional residues. *J. Mol. Biol.*, 344:1135–1146.

[2] Assenov, Y. et al. 2008. Computing topological parameters of biological networks. *Bioinformatics*, 24:282–284.

[3] Del Sol, A. et al. 2006. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol. Syst. Biol.*, 2:2006.0019.

[4] Kieffer, T.L. et al. 2007. Telaprevir and pegylated interferon-alpha-2a inhibit wild-type and resistant genotype 1 hepatitis C virus replication in patients. *Hepatology*, 46:631–639.

[5] Manns, M.P. et al. 2007. The way forward in HCV treatment—finding the right path. *Nat. Rev. Drug. Discov.*, 6:991–1000.

[6] Sarrazin, C. et al. 2007. Dynamic hepatitis C virus genotypic and phenotypic changes in patients treated with the protease inhibitor telaprevir. *Gastroenterology*, 132:1767–1777.

[7] Welsch, C. et al. 2008. Molecular basis of telaprevir resistance due to V36 and T54 mutations in the NS3-4A protease of HCV. *Genome Biol.*, in press.

[1]Max Planck Institute for Informatics, 66123 Saarbrücken, Germany. Email: `mario.albrecht@mpi-inf.mpg.de`
[2]Johann Wolfgang Goethe University Hospital, 60590 Frankfurt/Main, Germany. Email: `christophwelsch@gmx.net`

A Two-Stage Genome-Wide Association Study of Type 2 Diabetes Mellitus in a French Population

Johan Rung,[1] Ghislain Rocheleau,[1] Alexander Mazur,[1] Christian Dina,[2] Constantin Polychronakos,[3] Philippe Froguel,[2] Rob Sladek[1]

Until very recently, studies aiming at detecting association between genetic markers and a complex disease were mainly based on a candidate gene approach, where a small number of markers were selected for genotyping based on an *a priori* hypothesis of which genes would be most likely to be of importance to the disease. With new technologies for rapid and cost-effective large-scale genotyping, a number of genome-wide association studies (GWAS) have now been carried out for a number of diseases, scanning the entire genome for association without the need for any *a priori* hypothesis, often revealing unexpected risk loci that had not been identified in candidate gene studies. The results have been pouring in over the last year, with large and well-powered studies finding many risk loci for complex diseases. We published the first full genome-wide association scan for Type 2 Diabetes Mellitus (T2DM) last year [3], a case-control study using the Illumina BeadArray technology to study 392,935 SNPs across 1,363 French cases and control subjects, finding four new risk loci in addition to the previously discovered TCF7L2 association. Our study was followed by a number of other studies that confirmed two of our loci, HHEX and SLC30A8, in addition to discovering a number of additional loci [1, 2, 4, 5]. We have now finished the genotyping and analysis of Stage 2, covering the 15,688 most highly associated SNPs from Stage 1 in an additional 4,977 samples, and here report on our results.

It has been shown that a two-stage approach increases the power to detect risk variants, in particular computational analysis approaches combining data from the two stages. In this poster, we present the experimental and computational methods used for the selection and genotyping of SNPs for our Stage 2, the computational treatment of population stratification within our sample cohorts, and the analysis methods used for detecting association to disease across the two study stages. We also present a biological discussion of the found risk loci, in the light of other genome-wide association studies carried out by other groups.

References

[1] Saxena R., Voight B.F., Lyssenko V., et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316:1331–1336, 2007.

[2] Scott L.J., Mohlke K.L., Bonnycastle L.L., et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, 316:1341–1345, 2007.

[3] Sladek R., Rocheleau G., Rung J., et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445:881–885, 2007.

[4] Steinthorsdottir V., Thorleifsson G., Reynisdottir I., et al. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nature Gen.*, 39:770–775.

[5] Zeggini E., Weedon M.N., Lindgren C.M., et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, 316:1336–1341, 2007.

[1]McGill University and Genome Quebec Innovation Centre, 740 Doctor Penfield Ave, Montreal, QC, H3A 1A4, Canada. Email: `johan.rung@mail.mcgill.ca`

[2]CNRS 8090-Institute of Biology, Pasteur Institute, Lille 59019 Cedex, France.

[3]Dept. of Human Genetics, Faculty of Medicine, McGill University, Montreal H3H 1P3, Canada.

Detecting Significant Micro-Regions of DNA Aberration in High Density SNP Array Data

Gerard Wong,[1,2] Kylie Gorringe,[3] Ian Campbell,[3]
Izhak Haviv,[3‘] Christopher Leckie,[1,2] Adam Kowalczyk[2]

1 Introduction

High density Single Nucleotide Polymorphism (SNP) arrays provide a high resolution platform for the examination of DNA copy number aberrations and allelic imbalances in the human genome. Efficient and precise identification of such genetic aberrations and imbalances is critical in order to comprehend the biological processes that underlie the progression of diseases such as cancer, and also to offer an impetus for further medical diagnosis and the development of appropriate treatment. Significant challenges lie in analyzing high dimensional SNP data (up to 1.8 million probes on an Affymetrix SNP 6.0 Array) for which relatively few samples are available (typically in the order of hundreds). This challenge is compounded by the effects of experimental biases across samples and other forms of noise in the datasets. Our proposed approach involves the computation of a set of independent statistics across multiple samples to elucidate concordant regions of copy number change and loss of heterozygosity with p-values significantly smaller than the required Bonferroni correction threshold. The validation of our method to date has been achieved utilizing lung adenocarcinoma data from the Tumour Sequencing Project (TSP) [1] with results reported in [2]. We have identified a number of novel, statistically significant micro-regions of aberration in the datasets for which further biological verification is warranted.

2 Methods and Results

Figure 1 in [2] clearly illustrates substantial variation in smoothed copy number between samples, chromosomes and chromosome arms, albeit with some apparent large scale patterns consistent across the samples. There are various forms of biases in the data stemming from variations in experimental conditions, differences in hybridization on the microarrays and contamination of tumour samples with normal tissue. To counter these biases when detecting consensus regions of change, we have introduced two calibration methods.

1. **Rank calibration** replacing an original amplitude value by its cumulative probability density within a reference subset of all values for the sample (either for the whole genome, or for each chromosome or each chromosome arm). This is equivalent to the relative rank in the sorted values in the reference subset.

2. **Bipolar calibration** replacing an original value (after taking $\log_2$ and centering at 0) by –1 if it is $\leq \theta$, +1 if $\geq \theta$ and 0 otherwise, where is a selected threshold of 0.3 and 0.5.

Both calibration methods suppress outliers, control variance and allow for principled estimation of significance. The rank calibration converts measurements to the standard uniform distribution and naturally accounts for macro biases between samples, chromosomes and chromosome arms, respectively.

In our experiments we used calibrated measurements as described above as well as their smoothed (averaged) values with various flanking window sizes, specifically, $w = 0, 1, 2, 5, 10, 20$. Our basic filtering statistics were averages of calibrated and smoothed measurements across all samples; in the case of differentiation between binary phenotypes, differences between averages for each phenotype are calculated separately. The significance (p-value) was estimated as the probability of observing a value as extreme as the one observed for the convolution of filtering statistics over individual samples.

Figure 1 illustrates the results of our approach on 58 selected TSP lung adenocarcinoma samples [2] that passed various quality requirements. 18 statistics were computed for copy number (6 flanking window sizes matched with 1 rank and 2 bipolar calibrations of threshold 0.3 and 0.5. For LOH, 6

[1]Dept. of Comp. Sci. and Software Eng., Univ. of Melbourne, Australia. Email: gwong@csse.unimelb.edu.au
[2]NICTA Life Sciences, Victoria Research Laboratory, Australia. Email: adam.kowalczyk@nicta.com.au
[3]Peter MacCallum Cancer Centre, East Melbourne, Australia.

statistics were computed for each flanking window size. The plots display the $-\log_{10}$ ratio of p-values to the Bonferroni correction threshold $(2.38\times10^5)^{-1}$. The displayed values are truncated at 5 (upper bound) for concise representation. For example, at position 23656, the 'ranks' peak for flanking window, $w = 1$, corresponds to a p-value $< (2.38 \times 10^5)^{-1} \times 10^{-5} \approx 4.2 \times 10^{-11}$. Significant deletions on chromosome 13 are observed across 58 samples. In particular, notable alignments were observed between peaks derived from bipolar calibrations, rank calibration (for copy number) and LOH at probe position 23656 (13q12.2). The histograms show the distribution of $\log_2$ ratios over the 58 samples examined as well as for 263 TSP lung adenocarcinoma samples with quality requirements omitted. A distribution skewed towards negative values is evident in both histograms supporting the narrow region of deletion detected by our approach.

Our findings also reveal that through the application of base filtering statistics ($w = 0$), between 1K and 10K individual probes displayed p-values significantly below the Bonferroni-correction threshold of $(2.38 \times 10^5)^{-1}$. Hundreds of narrow peaks (in the order kilobases) were identified from consistent trends across multiple statistics. The selection of these peaks was assisted by data visualization with our developed software tools. A number of significant peaks were also found for the discrimination of several phenotypes in the dataset, such as that shown in Figure 2 (probe 1357) for males and females. These results demonstrate that our approach is complementary and mostly orthogonal to the analysis based on GLAD and GISTIC algorithms presented in [2].

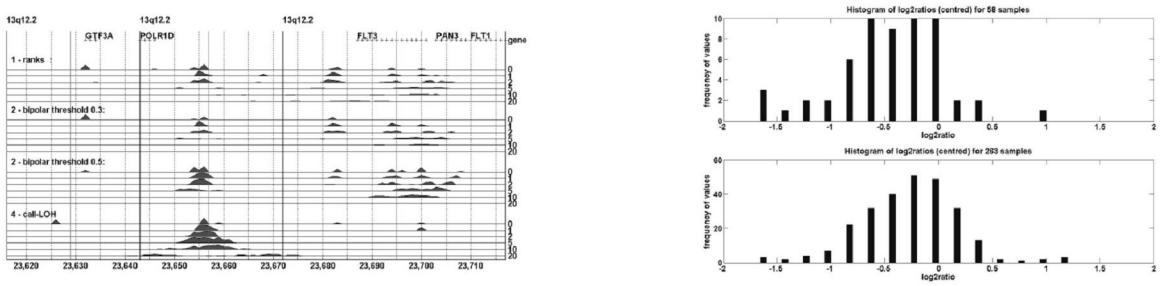


Figure 1: Example of significant micro-regions of deletion detected in lung adenocarcinoma dataset [1] and corresponding histograms of $\log_2$ ratios (centred at 0 for normal) at probe position 23656 across 58 and 263 TSP samples.

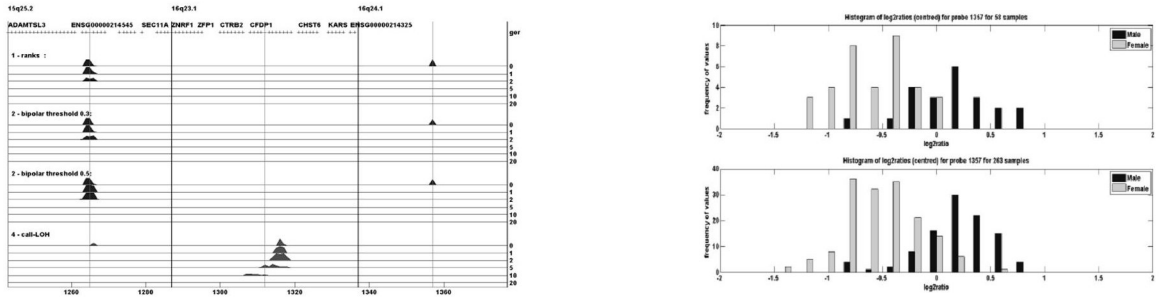


Figure 2: Example of significant differential peaks between males and females detected in lung adenocarcinoma dataset [1] and corresponding histograms of $\log_2$ ratios (centred at 0 for normal) at probe position 1357 across 58 and 263 TSP samples.

References

[1] Saxena R., Voight B.F., Lyssenko V., et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316:1331–1336, 2007.

[2] Scott L.J., Mohlke K.L., Bonnycastle L.L., et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, 316:1341–1345, 2007.

[3] Sladek R., Rocheleau G., Rung J., et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445:881–885, 2007.

[4] Steinthorsdottir V., Thorleifsson G., Reynisdottir I., et al. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nature Gen.*, 39:770–775.

[5] Zeggini E., Weedon M.N., Lindgren C.M., et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, 316:1336–1341, 2007.

Biomolecular Electrostatics: Beyond the Poisson-Boltzmann Centric view

Marc Delarue,[1] Patrice Koehl[2]

1 Introduction

Electrostatics interactions play a major role in the stabilization of biomolecules. As such, they remain a major focus of theoretical and computational studies in biophysics. Electrostatics in solution is strongly dependent on the nature of the solvent and on the ions it contains. While methods that treat the solvent and ions explicitly provide an accurate estimate of these interactions, they are usually computationally too demanding to study large macromolecular systems. Implicit solvent methods provide a viable alternative, especially those based on Poisson theory. The Poisson-Boltzmann equation (PBE) treats the system in a mean field approximation, providing reasonable estimates of electrostatics interactions in a solvent treated as continuum. In the first part of this paper, we review the theory behind the PBE, including recent improvement in which ions size and dipolar features of solvent molecules are taken into account explicitly. The PBE is a non linear second order differential equation with discontinuous coefficients, for which no analytical solution is available for large molecular systems. Many numerical solvers have been developed that solve a discretized version of the PBE on a mesh, either using finite difference, finite volume, finite elements or boundary element methods. Most of these methods have been optimized for the generic form of the PB equation, and as such cannot be applied directly to the modified PB equations, in particular those that include water dipole features explicitly. In the second part of the paper, we describe a new numerical method that solves all current forms of the PB equations.

2 Theory: Modified Poisson Boltzmann Equations

The Poisson-Boltzmann model is the most commonly used model to account for electrostatics interactions between charged objects. It assumes point-like charges immersed in a continuum dielectric medium and treats the system in a mean-field approximation. The medium is modeled by a homogeneous dielectric constant. In the presence of salt (one to one electrolyte), the Poisson Boltzmann equation is given by:

$$\nabla \cdot \left(\epsilon(r)\vec{\nabla}\phi(r)\right) - c(r)\kappa^2 \sinh\left(\frac{e_c\phi(r)}{k_B T}\right) = -4\pi e_c \sum_{i=1}^{M} q_i \delta(r - r_i) \tag{1}$$

where ϕ is the electrostatic potential, ϵ is the position-specific dielectric constant, κ is a coefficient that depends on the salt concentration, e_c is the charge of the electron, k_B is the Boltzmann constant, and T the temperature. The right hand side of the equation is the density of charges from the biomolecules.

It is important to note that the PB equation is based on many approximations. For the ions in the solution, it does not include ion size, nor does it account for ion-ion correlations. The medium itself is modeled by a homogeneous and isotropic dielectric constant; this assumption does not take into account the strong dielectric response of water molecules around charges. Modified Poisson-Boltzmann equations have been proposed to alleviate the problems related to these approximations. For example, Chu et al [1] introduced a size modified Poisson Boltzman (SMPB) equation to account for ion size. Abrashkin et al [2] proposed the dipolar Poisson Boltzmann (DPB) equation, which representes the medium by mobile dipoles, with orientable dipolar moment p. Recently, we proposed a generalized equation that combines the SMPB and the DPB equation into a generalized Poisson Boltzmann Langevin (GPBL) equation [3]. The general form of this equation is:

[1]Unité de Dynamique Structurale des Macromolécules, URA 2185 du C.N.R.S., Institut Pasteur, 25 rue du Dr Roux, 75015 Paris, France. Email: delarue@pasteur.fr

[2]Department of Computer Science and Genome Center, University of California, Davis, Davis, CA 95616, USA. Email : koehl@cs.ucdavis.edu

$$\begin{aligned}\frac{\epsilon_0}{4\pi}\Delta\Phi(\vec{r})(1+\frac{4\pi\lambda_{dip}\beta p_o^2 F_1(u)}{\epsilon_0 a^3 D(\Phi(\vec{r}))}) &+ \rho_f(\vec{r}) \\ &= \frac{2\lambda_{ion} ez\sinh(\beta ez\Phi(\vec{r}))}{a^3 D(\Phi(\vec{r}))}(1+\lambda_{dip}u^2 F_1(u)/D(\Phi(\vec{r}))) \\ &- \frac{\lambda_{dip}\beta^3 p_o^4 \vec{\nabla}\Phi(\vec{r}).(\vec{\nabla}\Phi(\vec{r}).\vec{\nabla})\vec{\nabla}\Phi(\vec{r})}{a^3 D(\Phi(\vec{r}))}(F_1^{'}(u)/u-\lambda_{dip}F_1(u)^2/D)\end{aligned}$$

with $u=\beta p_0|\vec{\nabla}(\Phi(\vec{r})|$, $D(\Phi(\vec{r}))=1+2\lambda_{ion}\cosh(\beta ez\Phi(\vec{r}))+\lambda_{dip}\frac{\sinh(\beta p_0|\vec{\nabla}\Phi(\vec{r})|)}{\beta p_o|\vec{\nabla}\Phi(\vec{r})|}$, $F_1(u)=\frac{1}{u}\left(\frac{u\cosh u-\sinh u}{u^2}\right)$, $\frac{F_1'(u)}{u}=\frac{1}{u}\frac{\partial F_1(u)}{\partial u}$. λ_{dip} and λ_{ion} are the fugacities of the dipoles and ions, respectively, and a is the common radius of the dipoles and ions. Note that the term $D(\Phi(\vec{r}))$ enforces steric avoidance.

3 Solving the Modified Poisson Boltzmann Equations

The Poisson Boltzmann Equation 1 is a second order nonlinear elliptic partial differential equation. Analytical solution of the PBE is only available for simple geometry such as spheres and cylinders [4, 5]. For the complex geometry of a biomolecule like a protein or a nucleic acid, analytical solutions are not available and the PBE must be solved using numerical methods [6]. Many such solvers have been developed [7, 8, 9].

The SMPB equation does not change the general structure of the PB equation, and as such can be solved using the same solver. For example, the SMPB equation was recently implemented in the package APBS.

The situation is different for the GPBL equation, as it include first order terms as well as a non linear Laplacian operator. We show in the paper that a modified Newton multigrid method can be used to solve the GPBL equation efficiently.

Acknowledgments. PK acknowledges support from the National Institute of Health under contract GM080399.

References

[1] V.B. Chu, Y. Bai, J. Lipfert, D. Herschlag, and S. Doniach. Evaluation of ion binding to DNA duplexes using a size-modified Poisson-Boltzmann theory. *Biophys. J.*, 93:3202–3209, 2007.

[2] A. Abrashkin, D. Andelman, and H. Orland. Dipolar Poisson-Boltzmann equation: Ions and dipoles close to charge interfaces. *Phys. Rev. Lett.*, 99:77801, 2007.

[3] C. Azuara, E. Lindahl, P. Koehl, H. Orland, and M. Delarue. Incorporating dipolar solvents with variable density in the Poisson-Boltzmann treatment of macromolecule electrostatics. *Nucl. Acids. Res.*, 34:W34–W42, 2006.

[4] C. Tanford and J. G. Kirkwood. Theory of protein titration curves. I. General equations for impenetrable spheres. *J. Am. Chem. Soc.*, 79:5333–5339, 1957.

[5] C.J. Benham. The cylindrical Poisson-Boltzmann equation. I. Transformations and general solutions. *J. Chem. Phys.*, 79:1969–1973, 1983.

[6] P. Knabner and L. Angermann. *Numerical Methods for Elliptic and Parabolic Partial Differential Equations*. Springer, 2003.

[7] M. Holst, N. A. Baker, and F. Wang. Adaptive multilevel finite element solution of the Poisson-Boltzmann equation I. Algorithms and examples. *J. Comp. Chem.*, 21:1319–1342, 2000.

[8] J. Liang and S. Subramaniam. Computation of molecular electrostatics with boundary element methods. *Biophys. J.*, 73:1830–1841, 1997.

[9] N. A. Baker, D. Sept, J. Simpson, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA*, 98:10037–10041, 2001.

Prediction and Analysis of Reliable Rearrangement Events in Mammalian Evolution

Hao Zhao,[1] Guillaume Bourque[2]

1 Introduction

The analysis of genome rearrangements can provide a whole-genome view on the evolution of related species. However, it is quite challenging to recover the true rearrangement evolutionary scenario of a set of contemporary genomes, even if the gene orders and the phylogenetic tree of these species are known. We recently proposed an *Efficient Method to Recover Ancestral Events* (`EMRAE`) to infer partial rearrangement scenarios consisting of only reliable ancestral events [1]. `EMRAE` can infer most types of rearrangement events: reversals, translocations, transpositions, fusions and fissions and thus it can be applied to the gene orders of both uni-chromosomal and multi-chromosomal genomes. We have shown that `EMRAE` has high sensitivities and specificities, which indicates that EMRAE can recover a significant part of the ancestral rearrangements with high reliability. Here we apply `EMRAE` to the gene orders of six mammalian genomes. Based on the simulations in [1], we are confident that `EMRAE`s predictions are highly reliable.

2 Methods and Results

2.1 EMRAE

Let T be the phylogenetic tree of a set G of genomes, and $e = (A, B)$ is an edge of T. The removal of e from T will partition the genome set G into two subsets S_A and S_B. To infer the ancestral events on the edge e, the idea of `EMRAE` was to identify the *conserved adjacencies* $CA(e, A)$ for S_A and the conserved adjacencies $CA(e, B)$ for S_B, and then combine them to trace back ancestral rearrangement events on the edge e; see [1] for details. `EMRAE` is applied to uni-chromosomal genomes and only infer reversals and transpositions in [1]. Based on the similar idea, we extend `EMRAE` so that it can also be applied to multi-chromosomal genomes and recover translocations, fusions and fissions.

2.2 Real Data

We apply `EMRAE` to the gene order data of six mammalian genomes: human (hg18), rat (rn4), mouse (mm9), dog (canFam2), chimp (panTro2) and rhesus (rheMac2). Using the same method as in [2] with a threshold 10 kb, we generate the orders of 3356 conserved blocks shared by the genomes. About 86% of the human sequence is covered by the identified blocks. Our predictions on the phylogenetic tree of the set of mammals are shown in Figure 1. See Figure 2 for an example of a predicted reversal on the human lineage illustrated by the alignment nets in the UCSC genome browser.

3 Future Work

Starting from the highly reliable predictions by `EMRAE` we will perform a further analysis on the sequence features at the breakpoint regions of these events. Specifically, we will study the pairing segmental duplications and other repeat families at the two breakpoint regions of each predicted reversal and translocation, since it has been revealed that these sequence features have a strong relationship with rearrangements. We hope that such an analysis will lead to new insight into the underlying mechanisms that shape the architecture of modern genomes.

So far `EMRAE` focuses on the recovery of rearrangement operations that only affect gene orders. It is also interesting to extend `EMRAE` to infer events that affect gene contents such like insertions and deletions.

[1]Genome Institute of Singapore. Email: `zhaoh1@gis.a-star.edu.sg`
[2]Genome Institute of Singapore. Email: `bourque@gis.a-star.edu.sg`

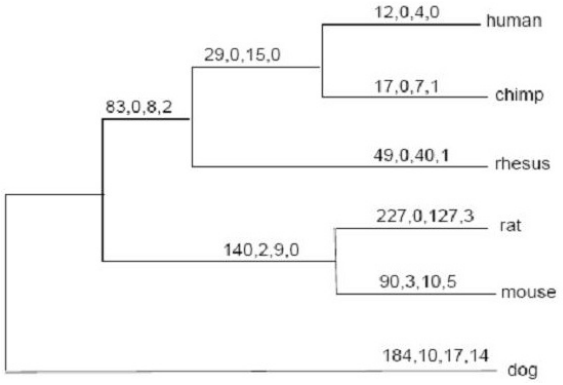


Figure 1: EMRAEs predictions on the phylogenetic tree of the six mammals. EMRAE recovers a total number of 1109 ancestral events, including 831 reversals, 15 translocations, 237 transpositions, and 26 fission/fusions. The 4 numbers on each edge represent the number of predicted reversals, translocations, transpositions, fusions/fissions on this edge, respectively.

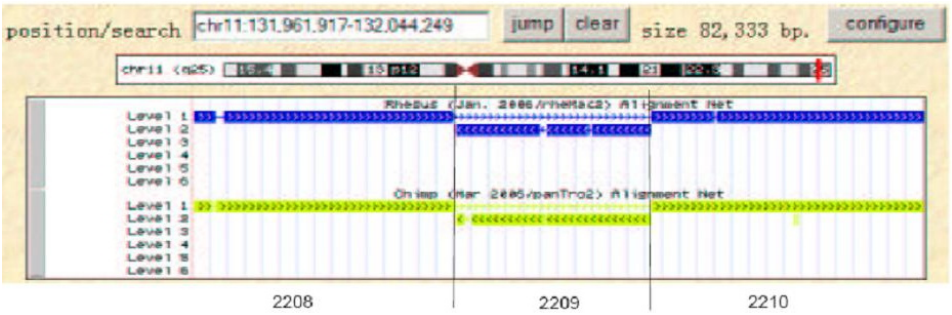


Figure 2: A reversal on the human lineage. The region shown is on the human chr11 and covered by three contiguous blocks 2208, 2209 and 2210. As illustrated by the alignment nets, both block 2208 and 2210 of three genomes have the same orientation, while block 2209 of rhesus and chimp has opposite orientation to that of human, which indicates that a reversal on the human branch flipped this block.

Acknowledgements. We would like to thank Jian Ma, who is from the Center for Biomolecular Science and Engineering, University of California Santa Cruz, for helping generate the gene order data.

References

[1] Zhao, H. and Bourque G. 2007. Recovering true rearrangement events on phylogenetic trees. In: *Fifth Annual RECOMB Satellite Workshop on Comparative Genomics*, pp.149–161.

[2] Ma, J., et al. 2006. Reconstructing contiguous regions in an ancestral genome. *Genome Research*, 16:1557–1565.

3D Structure Prediction of Camel Alpha-Lactalbumin

Maryam Nikousaleh,[1] Armin Madadkar Sobhani,[2] Bahram Goliaei[3]

1 Introduction

Prediction of protein 3D structure is one of the most challenging fields in bioinformatics. Among different theoretical methods of the protein 3D structure prediction, homology modeling as a more accurate technique in prediction of alpha-carbon coordinates, builds acceptable models of protein 3D structure in high similarity between target and template sequences [5]. Side chain conformations can be refined based on rotamer library of amino acid side chains [1].

Alpha-lactalbumin (alpha-LA) as a regulatory subunit of lactose synthase complex plays an important role in lactation process. Moreover, it is a calcium-binding protein and its apo form shapes a molten globule-like state in acidic pH and high temperature which make it a suitable model for studies of stability, folding and unfolding of calcium-binding proteins [4]. In this work, a homology model of camel alpha-lactalbumin is reported.

2 Methodology

MODELLER [5] has been regarded as one of the best homology modeling programs for prediction of protein 3D structure [6]. Therefore, the structure of camel alpha-LA was modeled by MODELLER version 9.2.

1B9O structure (Table 1), with 1.15Å structural resolution and 71% identity to the target sequence, was used as template. Structure prediction of camel alpha-LA was performed based on align2d of template and target. Initial model was evaluated with discrete optimized protein energy (DOPE) score calculation and PROCHECK program [2]. Refinement of structural parameters was carried out with MODELLER and SCWRL3 [5, 1]. After loop and side chains refinements, the final model was evaluated by PROCHECK program.

Structural stability of the model was confirmed by simulation using GROMACS 3.3.2 package [3]. Model was solvated in a constructed water box and its structural dynamics was simulated for 100 ps. Energy alterations and conformational changes were studied during simulation.

3 Results and Discussion

Final model has been shown in Fig. 1. According to the PROCHECK summary (Fig. 2), stereochemical parameters of our model were qualified in comparison with well-refined structures. DOPE energy plots of our model and template structures were very similar (Fig. 3). RMSD between model and templates is much lower than defined cut-off (e.g. 1.77Å against 3.5Å for superimposed alpha-carbons). Analysis of RMSD and structural energy plots of GROMACS show that conformational changes of the model are fairly fixed in equilibrium state (Fig 4, 5). These results confirm the accuracy of our model.

PDB ID	Title	Resolution	Identity	E-value
1B9O	Human alpha-LA structure in low temperature	1.15	71%	4.00E-45

Table 1: Features of template used for model building and results of alignment between template structure and target sequence.

[1]Laboratory of Biophysics and Molecular Biology, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran. Email: nikoosaleh@ibb.ut.ac.ir

[2]Department of Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran. Email: armin@ibb.ut.ac.ir

[3]Department of Biophysics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran. Email: goliaei@ibb.ut.ac.ir

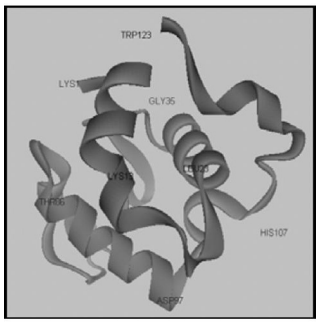


Figure 1: Built model for camel alpha-lactalbumin.

```
+----------<<< P R O C H E C K   S U M M A R Y >>>----------+
|                                                            |
| 1lac  1.1                                    123 residues  |
|                                                            |
| Ramachandran plot:   93.8% core    6.2% allow     .0% gener     .0% disall |
|                                                            |
| Gly & Pro Ramach:     0 labelled residues (out of   8)     |
| Chi1-chi2 plots:      1 labelled residues (out of  99)     |
|                                                            |
| Main-chain params:    6 better     0 inside      0 worse   |
| Side-chain params:    5 better     0 inside      0 worse   |
|                                                            |
| Residue properties: Max.deviation:     5.7          Bad contacts:    0 |
|                     Bond len/angle:    4.2    Morris et al class:  1  1  2 |
|                                                            |
| G-factors           Dihedrals:    .09  Covalent:   -.07    Overall:    .04 |
|                                                            |
| M/c bond lengths: 97.6% within limits   2.4% highlighted   |
| M/c bond angles:  84.2% within limits  15.8% highlighted   |
| Planar groups:   100.0% within limits    .0% highlighted   |
|                                                            |
+------------------------------------------------------------+
```

Figure 2: PROCHECK plot of the model evaluation.

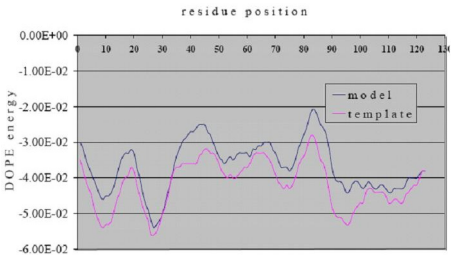


Figure 3: DOPE score plots of the model and template.

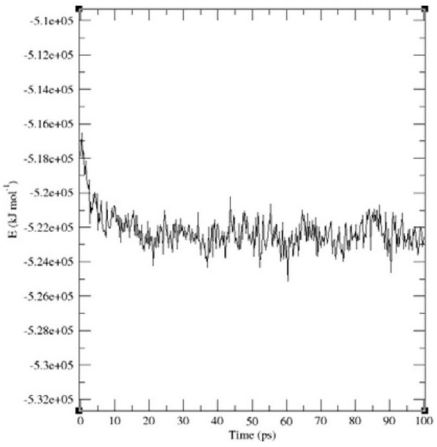


Figure 4: Energy alterations in natural dynamics of the model. Conformational changes of the model were simulated by GROMACS.

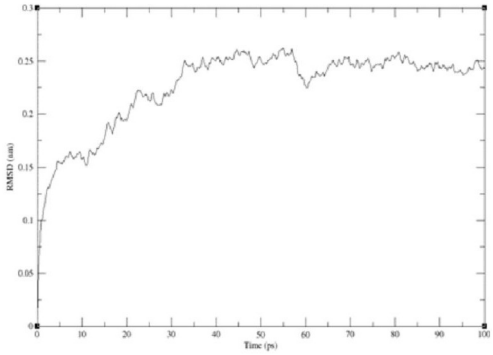


Figure 5: RMSD of the model conformations during simulation.

References

[1] Bower, M.J., Cohen, F.E., Dunbrack, R.L. 1997. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *J. Mol. Biol.*, 267:1268–1282.

[2] Laskowski R. A., MacArthur M. W., Moss D. S., Thornton J. M. 1993. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 26:283–291.

[3] Lindahl, E. Hess, B. and Spoel, D. 2001. GROMACS3.0: A package for molecular simulation and trajectory analysis. *J Mol Model*, 7:306–313.

[4] Permyakov, Eugene A. 2005. *Alpha-Lactalbumin*. USA: Vladimir N. Uversky.

[5] Sali, A., Blundell, T. L. 1993. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.*, 234:779–815.

[6] Wallner, B., Elofsson, A. 2005. All are not equal: A benchmark of different homology modeling programs. *Protein Science*, 14(5):1315-1327.

Comparative Analysis of Burkholderia Species Reveals an Association between Large-scale Genome Rearrangements and Fine-scale Nucleotide Variation in Prokaryotes

Chi Ho Lin,[1] Guillaume Bourque,[2] Patrick Tan[3]

1 Introduction

Large-scale genome gains and losses have long been recognized as important features of microbial evolution, however, the influence of rearrangement events such as translocations and inversions on functional diversity has been less explored. Specifically, it is unclear how rearrangements, by disrupting existing patterns of gene order and chromosomal organization, might specifically contribute to functional alterations in prokaryotic cellular pathways. Given the high frequency of such chromosomal rearrangements in bacterial families, exploring this issue is likely to be important and may further our understanding of gene-phenotype relationships in microbes.

2 Methods and Results

To investigate the relationship between genome rearrangements and nucleotide variation, we compared four closely-related members of the Gram-negative *Burkholderia* family (*B. pseudomallei*, *B. mallei*, *B. thailandensis* and *B. cenocepacia*) and identified a core set of 2590 orthologs present in all four species ("metagenes") using a reciprocal Blast strategy. Note that *Bp* is used as the reference genome in this study because of our prior interest in this organism as the causative agent of melioidosis [1].

The metagenes were unevenly distributed between the two *Burkholderia* chromosomes, and were organized into 255 synteny blocks whose relative order has been altered by a predicted minimum of 242 genome rearrangement events (see Figure 1) using the MGR algorithm [2]. Genes located within synteny blocks were significantly associated with common cellular functions compared to genes from different blocks, consistent with a non-random mode of chromosomal breakage biased against separating genes with common functionalities. We found that genes adjacent to rearrangement breakpoints ("boundary genes") exhibited higher levels of molecular divergence compared to genes interior to a synteny block (see Table 1) which we refer to as Boundary Element Associated Divergence (BEAD). This suggests a link between rearrangement breakpoints and local fine-scale genetic alterations.

To further validate the prevalence of BEAD, we revisited the set of metagenes that had experienced an inter-chromosomal translocation (or transposition) event in one of the *Burkholderia* lineages and tested if these translocated metagenes might be associated with increased divergence rates relative to the entire metagene population. Our results indicated that the translocated metagenes appear to be associated with elevated genetic divergence, an observation that is consistent with BEAD.

We also show that this phenomenon is detectable in both the *Pseudomonas* and the *Shigella* families (see Table 2), suggesting that this is a common phenomenon in prokaryotes.

3 Discussions

Our results suggest that balanced genome rearrangements may influence functional diversity and the development of novel microbial phenotypes by both the enhanced divergence of boundary genes, and by creating foci for the acquisition and deletion of species-specific genes.

[1]Genome Institute of Singapore, Singapore. Email: `linc@gis.a-star.edu.sg`
[2]Genome Institute of Singapore, Singapore. Email: `bourque@gis.a-star.edu.sg`
[3]Genome Institute of Singapore, Singapore; Duke-NUS Graduate Medical School, Singapore. Email: `tanbop@gis.a-star.edu.sg`

References

[1] Wiersinga, W. J., T. van der Poll, N. J. White, N. P. Day, and S. J. Peacock. 2006. Melioidosis: Insights into the pathogenicity of Burkholderia pseudomallei. *Nat Rev Microbiol*, 4:272–282.

[2] Bourque, G., and P. A. Pevzner. 2002. Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Res*, 12:26–36.

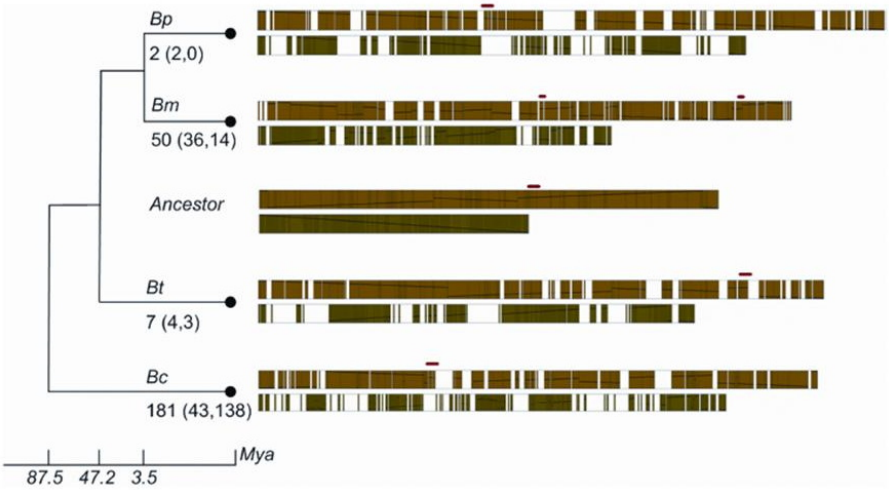


Figure 1: Whole genome synteny maps of *Bp, Bm, Bt, Bc* and of the predicted *Burkholderia* ancestor. Blocks and gaps are displayed proportionally to their actual size in all four species. Numbers within brackets indicate the number of reversals in Chr 1 and Chr 2 respectively while the number outside refers to the total number of reversals of both chromosomes.

	Bc		*Bt*		*Bm*	
Burkholderia	*Chr 1*	*Chr 2*	*Chr 1*	*Chr 2*	*Chr 1*	*Chr 2*
'Boundary genes' Percent Identity (%)	78.93	74.76	93.74	93.01	99.41	99.4
All-metagenes Percent Identity (%)	83.57	78.18	95.28	93.95	99.56	99.46
p-value (Wilcoxon)	2.28E-08	1.32E-09	4.01E-06	3.97E-06	1.11E-55	6.66E-16

Table 1: BEAD effect shown by comparing the percent identify of "Boundary metagenes" (Bmets) versus all-metagenes in *Burkholderia* using *Bp* as the reference.

	Pseudomonas			*Shigella*		
	Pseen	*Pp*	*Pfl*	*So*	*Sbo*	*Ss*
Bmets Percent Identity (%)	74.22	72.96	73.23	98.71	98.71	98.62
All-metagenes Percent Identity (%)	76.27	75.17	75.39	98.9	98.9	98.95
p-value (t-test)	0.0182	0.0299	0.0256	0.102	0.102	0.109
p-value (Wilcoxon)	1.33E-15	3.33E-16	7.77E-15	5.87E-03	3.39E-03	0.0354

Table 2: BEAD effect shown by comparing the percent identity of Bmets versus all-metagenes in *Pseudomonas* and *Shigella*.

EzArray: A Web-Based Highly Automated Affymetrix Expression Array Data Management and Analysis System

Yuerong Zhu,[1] Yuelin Zhu,[2] Wei Xu[3]

1 Introduction

Though microarray experiments are very popular in life science research, managing and analyzing microarray data are still challenging tasks for many biologists. Most microarray programs require users to have sophisticated knowledge of mathematics, statistics and computer skills for usage. With accumulating microarray data deposited in public databases, easy-to-use programs to re-analyze previously published microarray data are in high demand.

EzArray is a web-based Affymetrix expression array data management and analysis system for researchers who need to organize microarray data efficiently and get data analyzed instantly. EzArray organizes microarray data into projects that can be analyzed online with predefined or custom procedures. EzArray performs data preprocessing and detection of differentially expressed genes with statistical methods. All analysis procedures are optimized and highly automated so that even novice users with limited pre-knowledge of microarray data analysis can complete initial analysis quickly. Since all input files, analysis parameters, and executed scripts can be downloaded, EzArray provides maximum reproducibility for each analysis. In addition, EzArray integrates with Gene Expression Omnibus (GEO) and allows instantaneous re-analysis of published array data.

EzArray is a novel Affymetrix expression array data analysis and sharing system. EzArray provides easy-to-use tools for re-analyzing published microarray data and will help both novice and experienced users perform initial analysis of their microarray data from the location of data storage. We believe EzArray will be a useful system for facilities with microarray services and laboratories with multiple members involved in microarray data analysis. EzArray is available from `http://www.ezarray.com`.

2 Software and Files

To implement EzArray, we adopted the popular database and web application software bundle LAMP which refers to Linux operating system, Apache web server, MySQL database, PHP programming language. Selecting these technologies is mainly based on features such as low technical requirements for webmasters, programmers, and end users, open source, rapid application development, low total cost of ownership, and extremely large resources for free application source codes. In addition, we heavily incorporated Ajax (Asynchronous Javascript And XML) technologies to increase the systems interactivity, speed, functionality, and usability.

On EzArray server, PHP scripts deal with communication between users and the server, dynamically generate R scripts based on user input, execute R scripts in the background, and parse R output and present results to end users as HTML webpages. User information, data files, project information and analysis results are stored in database and server file system. EzArray comes with a web-based file management tool (My Files) and a request job management tool (Job List). On the client end, users logically follow these steps: register, logon, create or join a user group, create projects, import sample information and upload microarray data, submit analysis requests and browse results. The analysis tools (PreQ, ProS, and RepA) can be used in orders. Users can perform each type of analysis multiple times with modified parameters.

[1]BioInfoRx, Inc., Middleton, WI 53562, USA. Email: `ron@bioinforx.com`
[2]BioInfoRx, Inc., Middleton, WI 53562, USA. Email: `jack@bioinforx.com`
[3]Department of Oncology, University of Wisconsin-Madison, Madison, WI 53706, USA. Email: `wxu@oncology.wisc.edu`

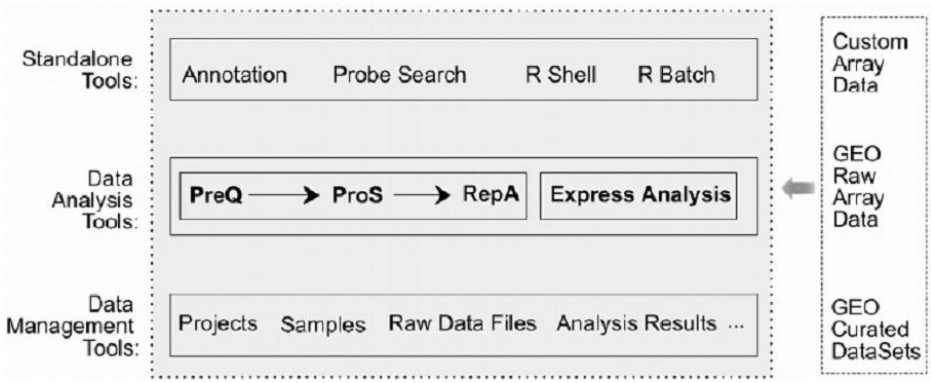


Figure 1: EzArray is an Affymetrix expression array data management and analysis system. EzArray can be used to manage and share data including projects, samples, raw array data files, and analysis results. EzArray includes three highly automated and seamlessly integrated data analysis programs named PreQ for data preprocessing and quality assessment, ProS for data processing and statistical testing, and RepA for report generating and gene annotation. Express Analysis is a one-step data analysis tool that covers all processing procedures in PreQ, ProS, and RepA. Microarray data can be from users experiments (Custom Array Data), published raw array data (deposited CEL supplementary files in GEO), or GEO curated DataSets (GDS records). In addition, a number of standalone tools have been included in EzArray, including tools for gene annotation, array probe search, R shell for interactive execution of R scripts, and R batch for batch execution of R scripts.

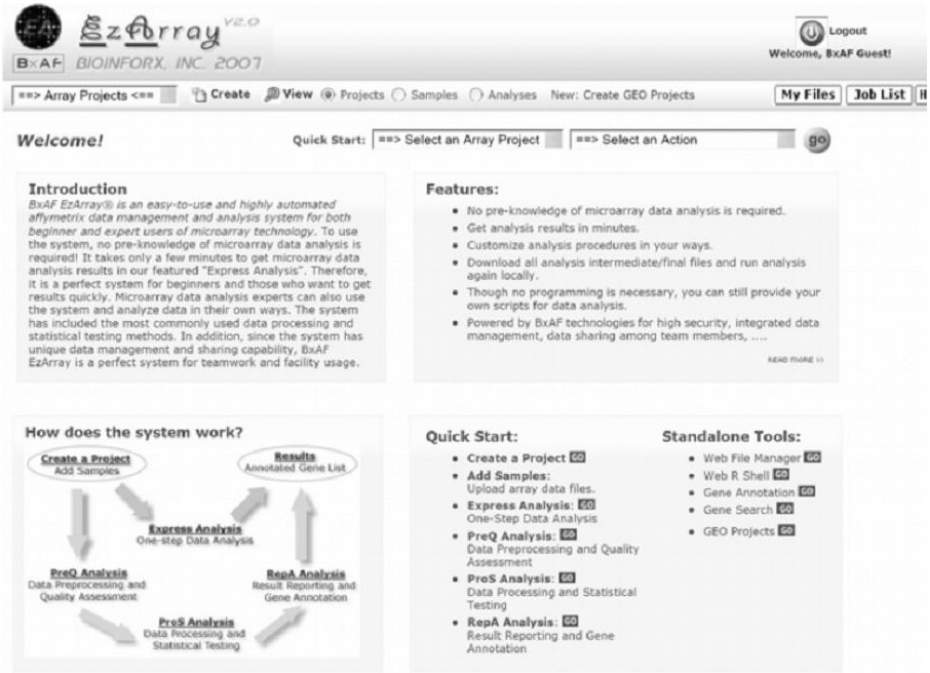


Figure 2: EzArray is a web-based system implemented with advanced web technologies. A screenshot of the EzArray homepage. The most important navigation tool in EzArray is the menu bar under the EzArray logo. However, users can also use the Quick Start pull-down menus, the hyper-linked diagram, or the Quick Start links to get started.

Estimating Signaling Networks Through Nested Effects Models

Holger Froehlich, Mark Fellmann, Annemarie Poustka
Holger Sueltmann, Tim Beissbarth[1]

1 Abstract

In the modern field of systems biology scientists aim to get insights into the architecture and behavior of complex cellular and genomic processes. An important task in this context is the detection of novel interdependencies between gene products. This insight into the molecular networks is an important step towards a better understanding of the functional aspects of a biological system. The advent of RNA interference techniques enables the selective silencing of biologically interesting genes in an efficient way. The combination of targeted interventions using the RNA interference technique with measuring effects on gene expression by DNA microarrays thus enables researchers to gain insights into the signal flow between proteins in a cell based on the observation of downstream effects. For example, in a signaling pathway that activates several transcription factors, blocking an upstream element of the pathway will affect all transcription factor targets, while perturbing one of the downstream transcription factor will only affect its targets, which are a subset of the genes effected by blocking the complete pathway. Markowetz et al. have proposed Nested Effect Models as a statistical framework for scoring networks hypotheses in a Bayesian manner.

We will show extensions of that framework that go in several directions: We show how prior assumptions on the network structure can be incorporated into the scoring scheme by defining appropriate prior distributions on the network structure as well as on hyperparameters. A new approach called module networks is introduced to scale up the original approach, which is limited to around 5 genes, to infer large scale networks. We compare several heuristic approaches for their performance in terms of sensitivity, specificity and speed. Instead of the data discretization step needed in the original framework, we propose the usage of a beta-uniform mixture distribution on the p-value profile, resulting from differential gene expression calculation, to quantify effects. Extensive simulations on artificial data and application of our module network approach to infer the signaling network between 13 genes in the ER-alpha pathway in human MCF-7 breast cancer cells show that our approach gives sensible results. Using a bootstrapping approach this reconstruction is found to be statistically stable.

2 Software

The code for the module network inference method is available in the latest version of the R-package nem, which can be obtained from the Bioconductor homepage.

References

[1] Froehlich H, Fellmann M, Sueltmann H, Poustka A, Beissbarth T. *Proc. German Conf. Bioinformatics* (GCB), 2007, 45–54.

[2] Tresch A, Beissbarth T, Sueltmann H, Kuner R, Poustka A, Buness A. *Journal of Computational Biology*, 2007,

[3] Froehlich H, Fellmann M, Sueltmann H, Poustka A, Beissbarth T. *BMC Bioinformatics*, 2007, 8(1):386.

[4] Froehlich H, Fellmann M, Sueltmann H, Poustka A, Beissbarth T. *Bioinformatics*, in press.

[1]DKFZ, Molecular Genome Analysis (B050), INF 580, 69120 Heidelberg, Germany. Email: `t.beissbarth@dkfz.de`

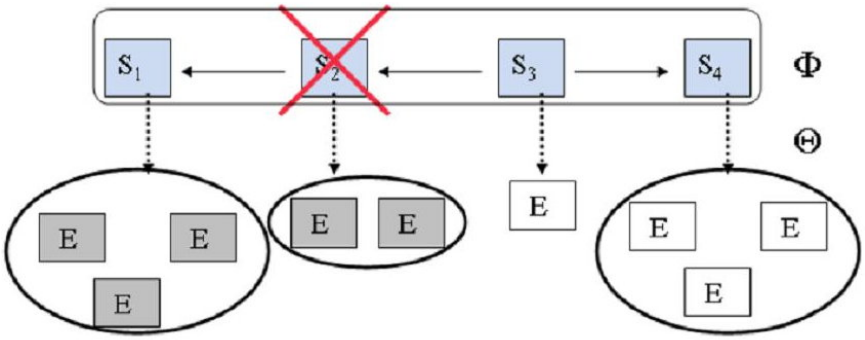


Figure 1: The combination of targeted interventions using the RNA interference technique with measuring effects on gene expression by DNA microarrays enables researchers to gain insights into the signal flow between proteins in a cell based on the observation of downstream effects. We distinguish between: a) S-genes, which are the silenced or signaling genes, for which the network (F) is to be determined. b) E-genes, i.e. genes for which an effect after an intervention is measured. The assignment of S-genes to E-genes is determined by the graph T. Displayed in the schema is an exemplary signaling network for which the expected effects after a single intervention are highlighted.

Statistics for Co-Occurrence of DNA Motifs

Utz J. Pape,[1,2] Martin Vingron[1]

1 Introduction

An important goal in computational biology is to decipher the transcriptional regulation of genes. Genes are regulated by transcription factors (TFs), which bind mainly upstream of the gene to the DNA. The TFs recognize TF-specific motifs called TF binding sites (TFBSs). By interaction with nearby TFs, they can initiate or inhibit transcription of the gene [2]. Detection of co-operativity between TFs is a first step to understand combinatorial transcriptional regulation. Such TFs are assumed to have exceptionally many TFBS approximate to each other. Thus, a significant number of co-occurrences of the corresponding DNA motifs is used to assess the strength of co-operativity.

Usually, Position Frequency Matrices (PFMs) are used as model for DNA motifs [6]. The methods to detect co-operativity based on co-occurrences can be divided into approaches relying on small distances between TFBSs (e.g. [7, 8]) and equivalently on high number of TFBSs in a small window (e.g. [1, 3, 4]). The significance of the co-operativity is either calculated under assuming position independence [3, 7, 8] or employing a randomization [1, 4]. The position independence of binding site occurrences is strongly violated for (self-)overlapping TFBS [5, 7]. The significance calculation based on randomization also encounters problems for similar PFMs, hence, the authors remove similar PFMs from the analysis [4]. Also, incorporating the complementary strand, introduces further dependencies and worsen the results.

We propose an accurate approximation for the significance calculation of the co-operativity of pairs of TFs circumventing the position independence assumption, incorporating similarity between PFMs, and including the complementary strand. We call two TFs to be co-operative if the corresponding DNA motifs co-occur exceptionally often. Two DNA motifs co-occur if both TFs have at least one occurrence in a specified window. Hence, we split the sequence into equal-sized non-overlapping windows covering the whole sequence. Next, we count the number of windows with a co-occurrence of the given pair of TFs. We can compute the overall significance for co-operativity based on the Poisson distribution. The rate corresponds to the probability of a window with a co-occurrence of the two DNA motifs in a random sequence (i.i.d.). Considering the overlap probabilities between the occurrences of the TFs, we capture the (self-)overlap of the PFMs and most of the dependencies introduced by the complementary strand. We call this a first-order approximation since we ignore dependencies between three or more positions. The accuracy of the results is shown by comparing the probability of the co-occurrence with a simulation based on random sequences, as well, as the performance of the co-operativity p-value in comparison to a simulation.

2 Results and Discussion

The comparison of the approaches is based on five artificial PFMs. Since one can anticipate that the overlapping structure between PFMs influences the result of the approximation, we include PFMs with/without selfand inter-repetive elements. The PFM 'nothing' with consensus $AAACAAACCCCC$ has no self-overlapping structure. The 'repeat' motif ($[AC]^5$) is strongly self-repetitive, while the 'palindrome' motif ($AA[C/G]^4TT$) tends to have each hit twice (one on each strand). The 'overlap' motif ($ACACGT$) overlaps with 'repeat'.

We randomly generate 100 sequences each with a length of 1,000,000 using an i.i.d. model. The sequences are annotated by the PFMs. Subsequently, we count the number of windows with co-occurrences for each pair of PFMs for different window sizes. Based on this number, we can compute the empirical probability for a co-occurrence. The top row of Figure 1 compares the empirical probability with the results of a model ignoring dependencies and the new approach for different window sizes. The results for the independence model are consistenly biased towards too high values except for the pair 'nothing' - 'overlap'. Only this pair neither contains a palindromic structure nor an overlap. The new approach always yields accurate results. Obviously, the palindromic and overlapping structure of DNA motifs can

[1]Dept. of Computational Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany. Email: utz.pape@molgen.mpg.de

[2]Dept. of Mathematics and Computer Science, Free University of Berlin, Berlin, Germany.

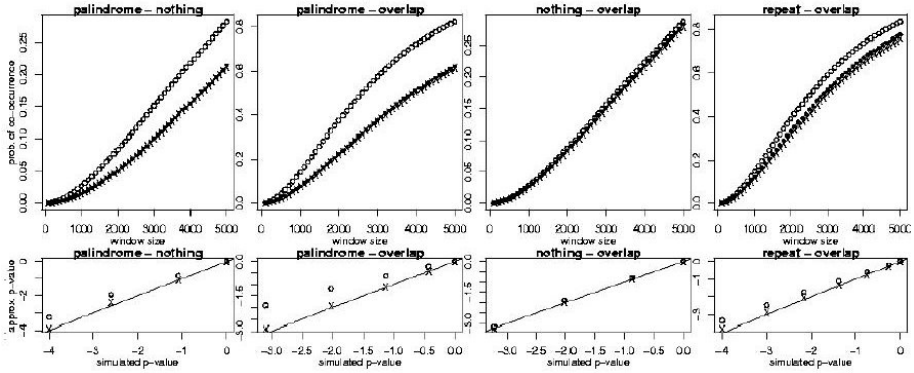


Figure 1: The upper row contains the comparison of the empirical probability (y-axis) for a co-occurrence event (small solid circles), an approach assuming independence (big circles), and the new approach (crosses) for different window sizes (x-axis). The lower row compares the corresponding logarithmic (to base 10) co-operativity p-values of the empirical distribution (x-axis) to the independence approach (circles) and the new apprach (crosses) for window of size 500.

have a strong influence on the probability of a co-occurrence. In contrast to the independence model, the new approach can deal with this.

The lower row of Figure 1 compares the p-values for the co-operativity between the simulated, the independence, and the new approach. The simulation is based on 10,000 sequences of length 10,000 which are annotated with the PFMs. Again, the number of windows with co-occurrences yields an empirical frequency. This is used to compute a p-value for the number of these windows based on a Poisson distribution. The independence model over-estimates the p-values for all pairs except nothing - overlap. This is not surprising since the probability for the co-occurence is strongly over-estimated for these pairs. In contrast, the new approach yields very accurate p-values for all pairs of TFs. Only the smallest p-values are slightly over-estimated. However, these small p-values just have minor support from the simulation. The different number of obtained p-values (points in the plot) are due to the different (co-)occurrence probabilities (see upper row of Figure 1) and the limited number of sequences and sequence length in the simulation.

The results show that (self-)overlap of PFMs influence the probability for co-occurrences. In contrast to standard approaches, our approximation sufficiently captures this bias by considering overlap probabilities. The new approach also incorporates the complementary strand. The major drawback is the simple background model (i.i.d. sequence). Extension to a Markov model might be complicated since many further dependencies are introduced. However, the i.i.d. sequence model has been widely and successfully used in computational biology. Furthermore, the proposed significance calculation will mainly be used for a filtering and ranking of co-operative TFs leading to experiments to confirm the hypotheses. In the future, we will extend the approach to multiple TFs to compute the significance of cis-regulatory modules. Incorporation of the observed frequency of a TF is planned, as well, as using overlapping windows.

References

[1] P. D. Bleser, B. Hooghe, D. Vlieghe, and F. van Roy. A distance difference matrix approach to identifying transcription factors that regulate differential gene expression. *Genome Biol*, 8(5):R83, 2007.

[2] J. W. Fickett. Coordinate positioning of mef2 and myogenin binding sites. *Gene*, 172(1):GC19–GC32, 1996.

[3] M. C. Frith, J. L. Spouge, U. Hansen, and Z. Weng. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res*, 30(14):3214–3224, 2002.

[4] S. Hannenhalli and S. Levy. Predicting transcription factor synergism. *Nucleic Acids Res*, 30(19):4278–4284, 2002.

[5] U. J. Pape, S. Rahmann, F. Sun, and M. Vingron. Compound Poisson approximation of number of occurrences of a Position Frequency Matrix (PFM) on both strands. *J. Comput Biol*, 2008, to appear.

[6] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res*, 10(9):2997–3012, 1982.

[7] A. Wagner. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, 15(10):776–784, 1999.

[8] W. Wasserman and J. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol*, 278(1):167–181, 1998.

Computational Simulations Suggest Transcription Factors AP-1 and NF-κB are Key Regulators of TLR3 Signaling

Mohamed Helmy,[1] Masaru Tomita,[1] Masa Tsuchiya,[1,*] Kumar Selvarajoo[1]

1 Introduction

Toll-like Receptor (TLR) 3 is an intracellular pattern recognition receptor involved in the innate response against viral infections. It recognizes double-stranded RNA (dsRNA) formed by most viruses during the duplication process. It is well studied that stimulation of TLR3 by dsRNA recruits adaptor molecule TRIF. This recruitment leads to activation of transcription factors IRF-3/7 and induction of proinflammatory cytokines (TNF-α and IL-6, etc) and interferons (CXCl10 and RANTES, etc). The dysregulation of these signaling processes plays a major role in the pathogenesis of viruses such as influenza A [1] and major illnesses such as autoimmune diseases [3], where the levels of proinflammatory cytokines and interferons often show significant increase. Thus, in order to understand and control proinflammatory responses, a systemic approach is required to identify optimal regulators of immune signaling pathways.

In this study, we developed a computational model of the TLR3 pathway and simulated the temporal activation dynamics of transcription factors IRF-3/7, AP-1 and NF-κB and the temporal induction of key proinflammatory cytokines TNF-α and IL-6. We particularly focused on AP-1 and NF-κB as they are known to play important role in immune response as well as in many other cellular processes such as apoptosis, cell differentiation and cell proliferation [5], however, their biological roles in TLR3 signaling have not been clearly understood. We compared our model simulations with similar experiments performed on murine macrophages [2]. To determine the optimal target(s) for abberated proinflammatory response regulation, we performed *in silico* Knock-outs (KO) and Knock-downs (Kd) simulations of all known signaling molecules in TLR3 pathway and found the double downregulation (but not abolishment) of AP-1 and NF-κB results in the most desired control for TNF-α and IL-6 in overt dsRNA response. Our results suggest that AP-1 and NF-κB, not just IRF-3/7, are also key transcription factors of TLR3 signaling.

2 Methods

We developed an *in silico* model utilizing signaling network originally obtained from the KEGG database and published experimental data (Fig. 1, [4]). Each reaction in our model was represented using mass-action kinetics with pulse perturbation given to TLR3 to represent the onset of signal transduction. *In silico* KOs were performed by setting the reaction upstream of the KO molecule to null, while Kds were generated by slowing down the Kd molecules upstream reaction kinetics. The details of modeling and parameters selection can be found in [6], the model details can be found in [4].

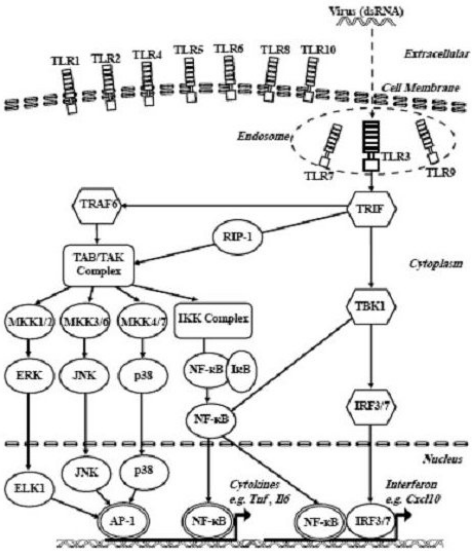


Figure 1: Schematic representation of TLR-3 signaling pathway adapted from [2].

[1]Institute for Advanced Biosciences, Keio University, Tsuruoka, 997-0017, Japan.
* Correspondence: {kumar, tsuchiya}@ttck.keio.ac.jp

3 Results and Discussion

In viral diseases and pro-inflammatory diseases, such as influenza A and osteoarthritis respectively, the induction of cytokines increases significantly [1, 3]. In order to control this increase we need to identify the molecule(s) that is (are) crucial for mediating the cytokines induction so that it (they) could be potentially targeted for therapeutic purposes. We therefore, performed several *in silico* KOs and Kds to investigate the effect on downstream TLR3 response; the activation of NF-κB and AP-1, and the induction of *Tnf*, *Il6* and *Cxcl10* mRNA. Among all the KOs simulations (data not shown), removing either NF-κB or AP-1 show significant effects in reducing the induction of *Tnf* and *Il6* (Fig. 2 C and D, WT, NF-κB KO and AP-1 KO), but they do not affect Cxcl10 levels (data not shown). This result indicates both NF-κB and AP-1 can be the optimal regulators of TLR3 signal induced pro-inflammatory cytokines.

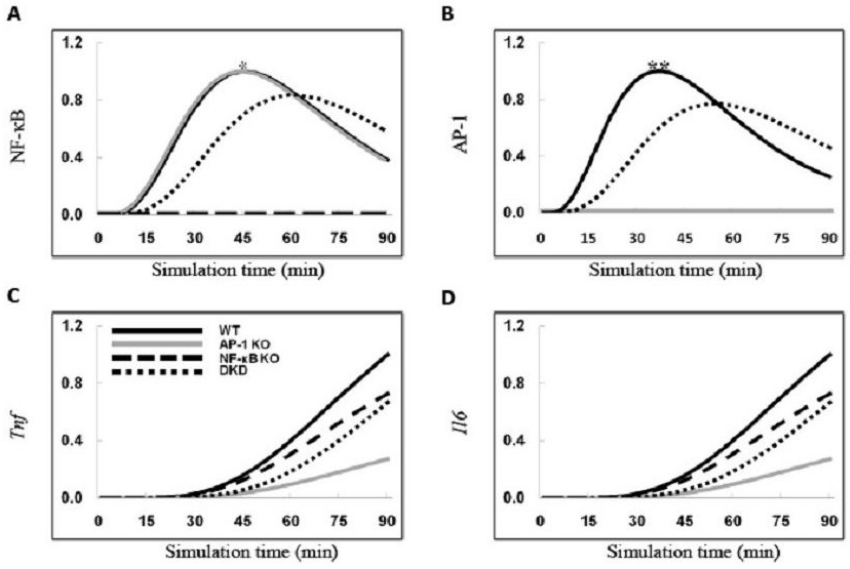


Figure 2: Simulation time course of A) NF-κB, B) AP-1, C) *Tnf* and D) *Il6*. Straight black line indicates wildtype (WT), dashed line indicates NF-κB Knock out (KO), gray line indicates AP-1 KO and dotted line indicates NF-κB/AP-1 Double Knock Down (DKD). The x-axis represents the simulation time in minutes and the y-axis represents the relative activation. (*) WT and AP-1 KO curves are overlapping. (**) WT and NF-κB KO curves are overlapping.

However, removing either NF-κB and AP-1 or both is likely detrimental to many other important cellular processes. Therefore, knocking out either of them may not be a viable option. We investigated, *in silico*, the possibility of producing similar results by reducing their overall reaction dynamics, rather than completely eliminating them. By doing so, Kds of NF-κB and AP-1, we were still able to obtain desirable outcome (Fig. 2 A, B, C and D, DKD). Thus, double targeting of NF-κB and AP-1 can be promising way of controlling cytokines increase in viral and pro-inflammatory diseases.

So far, it is well established that the main transcription factors of TLR3 pathways are IRF-3 and IRF-7. In this study, we have shown that NF-κB and AP-1 can also be key transcription factors for dsRNA response and should be collectively considered as possible target for pro-inflammatory control.

References

[1] Chan, M. Cheung, C. Chui, W. Tsao, S. Nicholls, J. Chan, Y. Chan, R. Long, H. Poon, L. Guan, Y. and Peiris, J. 2005. Proinflammatory cytokine responses induced by influenza A (H5N1) viruses in primary human alveolar and bronchial epithelial Cells. *Respiratory Research*, 6:135–147.

[2] Gohda, J. Matsumura, T. and Inoue, J. 2004. Cutting Edge: TNFR-Associated Factor (TRAF) 6 Is Essential for MyD88-Dependent Pathway but Not Toll/IL-1 Receptor Domain-Containing Adaptor-InducingIFN-(TRIF)-Dependent Pathway in TLR Signaling. *Journal of Immunology*, 173:2913–2917.

[3] Hana, H. Veronika, B. Zdenek, K. Marketa, P. Milan, A. and Ladislav, S. 2007. Increased level of cytokines and matrix metalloproteinases in osteoarthritic subchondral bone. *Cytokine*, 38:151–156.

[4] Helmy, M. Tomita, M. Tsychiya, M. and Selvarajoo, K. 2008. In silico analysis of Toll-Like receptor 3 pathways. In: *Proc. International Symposium on Computational Biology and Bioinformatics*, India (Bioinformatica Indica '08).

[5] Li, Q. and Verma, I. M. 2002. NF-kappaB regulation in the immune system. *Nat. Rev. Immunol.*, 2:725–734.

[6] Selvarajoo, K. 2006. Discovering differential activation machinery of the Toll-like receptor 4 signaling pathways in MyD88 knockouts. *FEBS Letters*, 580:1457–1464.

Modelling Metabolic Processes in Insulin-Secreting Pancreatic β-Cells

Lee Hazelwood,[1] John M. Hancock[2]

1 Introduction

The mouse is a widely used model of human disease. However mouse "models" of specific diseases do not recapitulate all the features of human diseases even when they carry a mutation that causes a disease in humans, because the genetic background of the model organism is different to that in humans. Because of this, there is a need to understand the consequences of individual genetic changes in as much detail as possible. Systems biology provides an opportunity to develop such an understanding because it allows us to put individual changes into a formally-defined systems context.

Type 2 diabetes mellitus (T2DM) affects at least 150 million individuals in the human population. Its frequency has increased rapidly over the past 20 years, most probably due to changes in diet. However there is strong evidence that predisposition to T2DM is genetic. We are collaborating with groups who are attempting to identify mouse and human mutations which predispose to T2DM or T2DM-like phenotypes.

Here we describe progress so far in developing a model of the processes taking place in pancreatic β-cells, the cells which are the primary secretors of insulin in the pancreas and therefore play a central role in diabetes. Pancreatic β-cells secrete insulin in response to the concentration of glucose circulating in the blood by using the glycolysis pathway as a sensor of glucose concentration. This is translated into changes in ATP concentration which lead to opening and closing of the K_{ATP} channel, subsequent opening/closing of the voltage-gated Ca^{2+} channel, and eventual secretion of insulin at appropriate levels.

2 The GSIS model

As part of an on-going development of a systematic model of insulin secretion by pancreatic β-cells, we developed a core metabolic model of the glucose-stimulated insulin secretion system (GSIS) [1]. This is a system of 44 ordinary differential equations which simulates the concentrations of 59 metabolites. The model incorporates aspects of a number of previous models of glycolysis, the TCA cycle, respiratory chain, NADH shuttles and the pyruvate cycle. Parameterization of the model was carried out using published values and systematic harmonization of the model parameters has yet to be carried out. The model incorporates three compartments: the cellular matrix and mitochondrial matrix and intermembrane space. It simulates the response of ATP to variations in glucose concentration in a realistic manner and shows oscillations in the concentrations of glycolysis metabolites and ATP broadly consistent with experimental observations. The model is encoded in SBML and executed using CVODE solvers [2, 3] and the Systems Biology Toolbox [4] in Matlab. The model is freely available to the academic community from the authors.

3 Extensions to the Core Model

A primary aim of the model is to understand the role of mutations in the protein*Nnt* (nicotinamide nucleotide transhydrogenase) in causing diabetes-like symptoms in laboratory mice [5], a model developed at MRC Harwell. *Nnt* is a component of the ROS (reactive oxygen species) detoxification system in mitochondria where it drives the reduction of hydrogen peroxide by glutathione - hydrogen peroxide is a by-product of electron transport which needs to be removed efficiently to avoid membrane decoupling and decreased levels of ATP production. The extended version of the model incorporates a model of glutathione-moderated ROS detoxification. This is less detailed than the rest of the model as some details of the reaction are still not known biologically, but we are collaborating with the Cox group at

[1]Bioinformatics Group, MRC Harwell, Harwell, Oxfordshire, UK. Email: l.hazelwood@har.mrc.ac.uk
[2]Bioinformatics Group, MRC Harwell, Harwell, Oxfordshire, UK. Email: j.hancock@har.mrc.ac.uk

MRC Harwell and other groups to develop this aspect of the model further. We are also extending our modelling to the downstream processes that respond to changes in ATP concentration: blocking of the K_{ATP} channels by elevated ATP concentration (in collaboration with F. Ashcroft, University of Oxford), membrane depolarization as a result of elevated intracellular K^+, opening of the voltage-gated Ca^{2+} channel and, ultimately secretion of insulin in response to higher intracellular Ca^{2+}.

Finally we are developing experimental collaborations that will allow more accurate parameterization of the core model. We are particularly interested in the conditions under which oscillation takes place in the glycolytic pathway and the aspects that control the frequencies of these oscillations.

Ultimately we aim to provide a model that will be useful to bench biologists in interpreting the results of their experimental work while also deriving a better understanding of the operation of a complex system critical to an important human disease.

Acknowledgments. This work was supported by ENFIN, a Network of Excellence funded by the European Commission within its FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health", contract number LSHG-CT-2005-518254.

References

[1] Jiang, N., Cox, R.D. and Hancock, J.M. 2007. A kinetic core model of the glucose-stimulated insulin secretion network of pancreatic -cells. *Mammalian Genome*, 18:508–520.

[2] Hindmarsh, A.C., Brown, P.N., Grant, K.E., Lee, S.L., Serban, R. et al. 2005. SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers. *ACM Transactions on Mathematical Software*, 31:363–396.

[3] Serban, R. Hindmarsh, A.C. 2005. CVODES: the Sensitivity-Enabled ODE Solver in SUNDIALS. In: *Proc. IDETC/CIE 2005*, Long Beach: ASME. pp. DETC2005–85597.

[4] Schmidt, H., Jirstrand, M. 2006. Systems Biology Toolbox for MATLAB: A computational platform for research in systems biology. *Bioinformatics*, 22:514–515.

[5] Freeman, H.C., Hugill, A., Dear, N.T., Ashcroft, F.M., Cox, R.D. 2006. Deletion of nicotinamide nucleotide transhydrogenase: A new quantitive trait locus accounting for glucose intolerance in C57BL/6J mice. *Diabetes*, 55:2153–2156.

Unsupervised Joint Analysis of ArrayCGH, Gene Expression Data and Supplementary Features

Christine Steinhoff,[1] Matteo Pardo,[2] Martin Vingron[3]

1 Introduction

The development of high throughput gene profiling methods, such as comparative genomic hybridization (CGH) and gene expression microarrays, enables for studying specific disease patterns in parallel.

The underlying assumption for studying both, genomic aberrations and gene expression is that genomic aberrations might effect gene expression either directly or indirectly. In cancer research, in particular, there have been a number of attempts to improve cancer subtype classification or study the relationship between chromosomal region and expression aberrations.

Most studies apply the following intuitive procedure for the analysis of aCGH and expression data: First determine regions with copy number aberrations (possibly tissue- or patients-specific) and then look for differentially expressed (onco)genes inside these regions [2, 5]. There is a natural reason for integrating results rather than data: strong heterogeneity does not allow sensible alignments of the source data. Still, integrative approaches where data are fused before their analysis- are preferable. Only recently, few integrative methods have been published [1, 6]. Nevertheless, these approaches do not integrate covariate data like tumor grading, staging, age, mutation status and other disease features. These features are frequently available and of interest for an integrative analysis. We address these two problems, namely jointly analyzing different data sources and integrating supplementary categorical data. Furthermore, our approach can easily be applied to diverse data sources, even more than two, with and without supplementary patients' information.

2 Methods

We established a new data analysis pipeline for joint visualization of microarray expression and arrayCGH data (aCGH), and the corresponding categorical patients' information. This pipeline comprises four parts: (a) data discretization, (b) binary mapping, (c) gene filtering, (d) multiple correspondence analysis. All computational analysis steps are programmed using R and Bioconductor [3, 4]. The first two steps transform data to a common binary format, a necessary step for jointly analyzing them.

(a) We propose three different approaches for the discretization of expression data: Probability of Expression, POE [9], ordinary fold change and DNAcopy [8, 11]. The different discretization procedures each focus on a different biological objective. For arrayCGH we use standard discretization with DNAcopy [8, 11].

(b) Discretized expression and arrayCGH data, and categorical supplementary data are mapped into a binary space by transforming each of the three data matrices to its corresponding indicator matrix.

(c) For many applications it is customary to remove noise and redundancy from omics data by reducing the number of features (genes). We considered variance filtering, expression-aCGH correlation filtering and PCA loading on the first two principal components.

(d) In the last step, we apply a method based on correspondence analysis, namely multivariate correspondence analysis with supplementary variables (MCASV) [7]. MCASV has been applied in the context of social sciences but to our knowledge has not been used in the context of biological high throughput data analysis. Features (expression and aCGH) and covariates (patients' information) are transformed into a common space. Vicinity between features and covariates can then be visualized and quantified. We e.g. determine genes that are correlated with covariates, possibly for interesting subsets of patients. In MCASV vicinity is measured by the angle intercurring between covariate and feature.

[1]Max Planck Institute for Molecular Genetics, Berlin, Germany. Email: christine.steinhoff@molgen.mpg.de
[2]SENSOR Laboratory, CNR-INFM, Brescia, Italy. Email: pardo@ing.unibs.it
[3]Max Planck Institute for Molecular Genetics, Berlin, Germany. Email: martin.vingron@molgen.mpg.de

3 Application and Results

We applied our approach to a published dataset on breast cancer. Pollack et al. [10] studied genomic DNA copy number alterations and mRNA levels in primary human breast tumors. We used the data preprocessed as described in [1] which results in a total of 6094 genes. The patients variables comprise tumor stage (stages 1 to 4), tumor grade (grades 1 to 3), node status (positive (+) or negative (-)), histology (ductal), ER status (positive (+) or negative (-)) and p53 status (wild type (wt) or mutant).

In the figure we show the plot obtained for all patients with the pipeline settings: POE for the discretization of expression values; filtering using PCA loadings. Gray points represent genes, while selected covariate values are represented by crosses and corresponding names.

As an example, we extracted those genes relating to 'Tumor stage 4' by taking genes which are in a 10 degree cone around 'Tumor stage 4' (circa 600). A Gene Ontology analysis of these genes showed significant enrichment of biological processes of anatomical structure development, cell growth, regulation of growth, defense response, responses to chemical and external stimulus, which all relate to cancer.

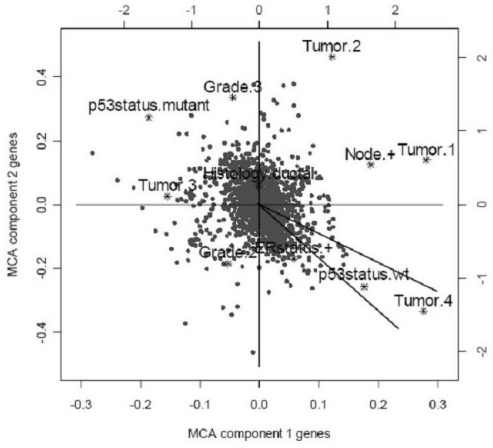


Figure 1: Output from joint analysis of expression and aCGH data with supplementary patients' information. Lower x and left y axis show first and second MCA component axis of the decomposition of gene states Burt matrix resp (black filled dots in the plot). Upper x axis and right y axis show first and second MCA component of covariate matrix decomposition (stars with covariates' names displayed above each star).

References

[1] Berger, J.A., et al., 2006. Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. *IEEE/ACM Trans Comput Biol Bioinform*, 3(1):2–16.

[2] Garraway, L.A. and W.R. Sellers, 2006. From integrated genomics to tumor lineage dependency. *Cancer Res*, 66(5):2506–2508.

[3] Gentleman, R., et al., 2004. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80.

[4] Ihaka, R. and R. Gentleman, 1996. R: A language for data analysis and graphics. *J. Comput. Graph. Stat.*, 5:299–314.

[5] Jacobs, S., et al., 2007. Genome-wide, high-resolution detection of copy number, loss of heterozygosity, and genotypes from formalin-fixed, paraffin-embedded tumor tissue using microarrays. *Cancer Res*, 67(6):2544–2551.

[6] Jeffery, I.B., et al., 2007. Integrating transcription factor binding site information with gene expression datasets. *Bioinformatics*, 23(3):298–305.

[7] Nenadic, O. and M. Greenacre, 2006, Multiple Correspondence Analysis and Related Methods. *Statistics in the Social and Behavioral Sciences Series*, London: Chapman & Hall/CRC. 523–552.

[8] Olshen, A.B., et al., 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572.

[9] Parmigiani, G., et al., 2002. A statistical framework for expression-based molecular classification in cancer. *JRSS*, 64:717–736.

[10] Pollack, J.R., et al., 2002. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci USA*, 99(20):12963–12968.

[11] Venkatraman, E.S. and A.B. Olshen, 2007. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663.

A Procedure to Identify MicroRNA Gene Targets in Human Kidney Cancer

H. Liu,[1] G. Alexe,[2] D. Juan,[3] T. Antes,[3] C. Delisi,[4] L. Liou,[3]
S. Ganesan,[5,*] G. Bhanot[1,5,6,*]

1 Introduction

MicroRNAs (miRNAs) are a class of naturally occurring, noncoding RNAs that regulate protein expression by targeting protein coding mRNA. It has been suggested that some miRNAs behave like oncogenes or tumor suppressor genes by regulating genes involved in biological functions such as celluar differentiation, development and apoptosis. Present computational predictions of gene targets of miRNAs are primarily based on identification of complements of miRNA sequences in the conserved 3'UTR region of genes and free energies of RNA-RNA duplexes [3, 4, 7]. These methods suffer from a high false positive rate and their predictions are not algorithm independent. Here, we demonstrate a direct method to identify gene targets of miRNA which correlates expression levels of miRNA and mRNA in matched normal kidney (NK) and renal cell carcinoma (RCC) samples. We identify candidate genes as those whose expression levels are highly (anti)-correlated with miRNAs and which are differentially expressed in tumor tissue compared to normal tissue. A gene enrichment analysis of the identified genes reveals many known RCC gene markers and biological pathways and directly pinpoints the miRNAs that regulate them.

2 Data and Method

Eight clear-cell RCC tissue specimens and surrounding NK tissue were collected from patients at Boston Medical Center and Cleveland Clinic and processed by standard methods. miRNA expression profiling was performed using real-time PCR in a 384-well format and normalized to "housekeeping" miRNAs, identified as those most unchanged across normal and tumor samples. The mRNA expression levels for the same samples were measured by hybridizing extracted RNA to Affymetrix HG-U133 Plus 2.0 arrays.

To identify mRNA targets of miRNAs in RCC, we used the hypothesis that "the expressions of miRNA and their target gene mRNAs should co-vary when averaged over matched samples". The most direct regulation corresponds to an anti-correlation between miRNAs levels and corresponding target mRNAs. Hence, the procedure we followed identified up/down regulated miRNAs in tumor samples relative to normal samples and then searched for putative mRNA targets that were down/up regulated respectively, also in tumor vs normal samples. We used the Pearson correlation coefficient at 1% significance on the 8 matched RCC/NK samples and the permutation test to evaluate the robustness of the correlations. Putative target mRNA were obtained from the TargetScan database Release 4.1 [4] (http://www.targetscan.org) which uses an algorithm that identifies regulatory targets of mammalian miRNAs by looking for conserved sites matching the seed region of each miRNA. We only considered miRNA families which are highly conserved across human, mouse, rat, dog and chicken.

Figure 1 is a flowchart of our procedure. We have retained for future analysis and experimental validation targets identified by our procedure but not captured in the database. Our method is based on experimental measurement and does not suffer from ambiguities of interpretation and modeling which plague sequence matching methods. It can be easily extended to other tumor types.

3 Preliminary Results and Discussion

We identified several known molecular markers of RCC. Expression levels of VEGF, a well-known protein highly expressed in RCC [8], were significantly correlated with mir-200bc/429 (permutation p-value $p = 0.0013$). 70 other genes were also identified as targets of this miRNA family, including oncogene APC and

[1]BioMaPS Institute, Rutgers University, Piscataway, NJ 08854.
[2]The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142.
[3]Boston University School of Medicine, 715 Albany St, Boston, MA 02118.
[4]Bioinformatics Program, 24 Cummington Street, Boston University, Boston, MA 02215.
[5]Cancer Institute of New Jersey, 195 Little Albany Street, New Brunswick, NJ 08093.
[6]Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ 08540.
* Joint Senior Authors. G.B. Email: gyanbhanot@gmail.com

TNFRSF6, growth factor GRB10 and VEGFC. More than 10 of these genes are located on Chromosome 5q (enrichment p-value=0.0038) which contains [1] two amplification regions in RCC: near 5q22 with APC, PJA2, SEMA6A, PRRC1 and UBE2B, and near 5q31 with CSNK1A1, CANX, CLK4 and YIPF5. In addition, we have identified 23 other miRNAs down-regulated in RCC with the following correlations: the EGFR gene, an important marker of RCC was correlated with mir-135 ($p = 0.007$); LOX correlated with mir-149 ($p = 0.009$) and several other oncogenes, for example, ECT2 (epithelial cell transforming sequence 2 oncogene) and RAP2B (a member of RAS oncogene family), correlating with the mir-204/211 family. These targeted genes are enriched in cell migration, extracellular matrix (ECM) organization and biogenesis and the ECM-receptor interaction pathway.

About 40 mRNAs were found to be suppressed in tumors by 9 up-regulated miRNAs. These include several tumor suppressor genes such as VHL, a gene known to be mutated or inactivated in $> 50\%$ RCC cases [6], correlated with mir-224 ($p = 0.0035$). In normal kidney, VHL inhibits VEGF and other hypoxia-inducible angiogenesis genes. Its loss leads to a microenvironment favorable for epithelial-cell proliferation and increases blood supply to the tumor [2]. Our observation of the control of VHL and other hypoxia-inducible genes with miRNA differentially expressed in RCC implies a role for miRNAs regulation of the hypoxia signaling pathway. Other identified targets of mir-224 include ERBB4 ($p = 0.0046$), a member of EGFR family reported strongly down-regulated in RCC and a potential tumor suppressor [9]. SFRP1, a negative regulator of the Wnt signaling pathway, is correlated with mir-34a ($p = 0.005$). Loss of SFRP1 expression is seen in a majority of RCC patients [5]. In summary, our results demonstrate that our procedure is an accurate and practical method to identify the targets of miRNAs in tumor development.

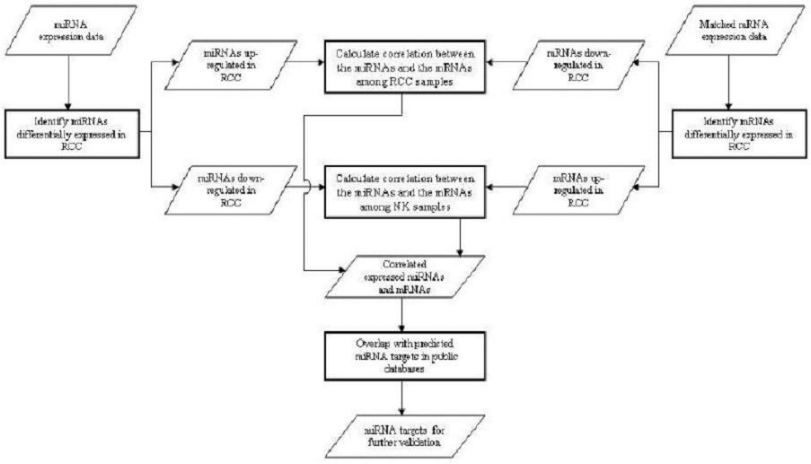


Figure 1: A flowchart of our proposed procedure to identify robust miRNA targets in RCC development. Differentially expressed miRNAs and mRNAs are obtained using computational and statistical methods from several expression experiments [Alexe et al. unpublished data].

References

[1] Bugert P, Knobloch RV, Kovacs G (1998) Duplication of two distinct regions on chromosome 5Q in non-papillary renal-cell carcinomas. *Int. J. Cancer*, 76:337–340.

[2] Cohen HT, McGovern FJ (2005). Medical progress: renal-cell carcinoma. *N Engl J Med*, 353:2477–2490.

[3] Enright AJ, John B, Gaul U, Tuschl T, Sander C and Marks DS (2003). MicroRNA targets in Drosophila. *Genome Biol*, 5:R1.

[4] Grimson A, Farh KKH, Johnston HK, Garrett-Engele P, Lim LP, Bartel DP (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, 27:91–105.

[5] Gumz ML, Zou H, Kreinest PA, Childs AC, Belmonte LS, LeGrand SN, Wu KJ, Luxon BA, Sinha M, Parker AS, Sun LZ, Ahlquist DA, Wood CG, Copland JA (2007). Secreted frizzled-related protein 1 loss contributes to tumor phenotype of clear cell renal cell carcinoma. *Clin Cancer Res*, 13(16):4740–4749.

[6] Kaelin WG Jr. (2004). The Von Hippel-Lindau Tumor Suppressor Gene and Kidney Cancer. *Clin. Cancer Res*, 10:6290S–6295S.

[7] LinksKrek A, Grn D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N (2003). Combinatorial microRNA target predictions. *Nat Genet*, 37(5):495–500.

[8] Takahashi A, Sasaki H, Kim SJ, Tobisu K, Kakizoe T, Tsukamoto T, Kumamoto Y, Sugimura T, Terada M (1994) Markedly increased amounts of messenger RNAs for vascular endothelial growth factor and placenta growth factor in renal cell carcinoma associated with angiogenesis. *Cancer Res*, 54:4233–4237.

[9] Thomasson M, Hedman H, Junttila TT, Elenius K, Ljungberg B, Henriksson R (2004). ErbB4 is downregulated in renal cell carcinoma - a quantitative RT-PCR and immunohistochemical analysis of the epidermal growth factor receptor family. *Acta Oncol*, 43:453–459.

An Integrated Probabilistic Approach for Gene Function Prediction Using Multiple Sources of High-Throughput Data

Trupti Joshi,[1] Chao Zhang,[1] Ning Lin,[1] Dong Xu[1]

1 Introduction

Determination of gene function is one of the most important problems in the post-genomic era. Characterizing gene function using large-scale biological data in an automated fashion can provide valuable hypotheses for biological studies. The key to accurate function prediction lies in the integration of these data, which is a challenging subject. Different approaches towards using single or integrating different data types have been developed over the years [1, 3, 5]. Nevertheless, most of the methods are developed for research purpose only and few approaches have been implemented as publicly available software. Considering the complexity of handling different types of high-throughput data and long computing time, most of the applications are still web-based containing precomputed gene function predictions, and are not readily extendible to predictions for user-supplied data sources

2 Results

To address this issue, we have developed GeneFAS (Gene Function Annotation System) [4], a gene function prediction method which utilizes all types of high-throughput data including microarray, SAGE, inparanoid, phylogenetic, protein-protein interactions and protein domain information. GeneFAS uses Gene Ontology [7] annotations for index level comparison of functional similarities. The method quantifies the relationship between functional similarity and underlying high-throughput data, and codes the relationship into a 'neighborhood graph', where each node represents one gene and each edge shows the Bayesian probability of function similarity between two genes [4, 2].

For example, consider a protein X with unknown function has associations with proteins U, V and W with known functions. With the assumption that F_i, $i = 1, 2, ..., n$, represents a collection of all the functions that proteins U, V and W have, a likelihood score for protein X to have function F_i, $G(F_i|X)$, is defined as:

$$G(F_i|X) = 1 - (1 - P'(S_l|M))(1 - P'(S_l|B))(1 - P'(S_l|C)) \quad (1)$$

where S_1 represents the event that two proteins have the same function, F_i, whose GO INDICES share l levels. For a given F_i, the probabilities of sharing the same function from various data type, $P'(S_l|M)$, $P'(S_l|B)$, andi $P'(S_l|C)$ are calculated based on the Bayesian probabilities of interaction pairs defined by gene expression correlation coefficient (M), protein-protein interaction (B), phylogenetic (C), etc. respectively. In each type of high-throughput data, a protein with unknown function might have multiple interaction partners with function F_i. Suppose that there are n_M, n_B,and n_C interaction partners with function F_i in the three types of high-throughput data, respectively. $P'(S_l|M)$, $P'(S_l|B)$, and $P'(S_l|C)$ in equation (1) are calculated as:

$$P'(S_l|M) = 1 - \prod(1 - P_j(S_l|M)),\ j = 1, 2, \ldots, n_M \quad (2)$$

$$P'(S_l|B) = 1 - \prod(1 - P_j(S_l|B)),\ j = 1, 2, \ldots, n_B \quad (3)$$

$$P'(S_l|C) = 1 - \prod(1 - P_j(S_l|C)),\ j = 1, 2, \ldots, n_C \quad (4)$$

We applied GeneFAS to predict mouse gene functions using the MouseFunc competition datasets [6], as outlined in Figure 1. The GO terms are grouped in 12 evaluation categories for evaluation purposes and these categories are corresponding to all combinations of 3 GO Ontologies—Biological Process, Molecular Function, and Cellular Component—with 4 ranges of "functional specificity" which is defined as the number of genes in the training set assigned to a particular GO term, i.e. [3–10], [11–30], [31–100] and [101–300]. A variety of performance measures like area under the ROC curve (AUC), precision at

[1]Digital Biology Laboratory, Computer Science Department and Christopher S. Bond Life Sciences Center, 1201 East Rollins Road, University of Missouri-Columbia, Columbia, MO 65211-2060, USA. This work was supported by USDA/CSREES-2004-25604-14708 and NSF/ITR-IIS-0407204.

different recall % are applied. These measures are applied to each GO term individually, and mean performance values are calculated for 12 categories of GO terms. The prediction results (Table 1) were evaluated against two sets of genes: 1) test set of held-out genes, and 2) novel set of genes that had been newly annotated to a GO term in the training set during the eight months since downloading of the version of MGI GO annotation. Performance evaluation using single datasets for function prediction shows that the Pfam and InterPro are the most important data sources for function prediction for all three GO Ontologies. Microarray Su dataset also contributes very useful information and protein-protein interactions data seems to contribute the least when used singly. Also, "Maryland-bridge coefficient" performs the best in all subcategories when compared to "Pearson's Correlation Coefficient" and "Jaccard coefficient". GeneFAS has a robust performance and gave good results in both the novel and test sets.

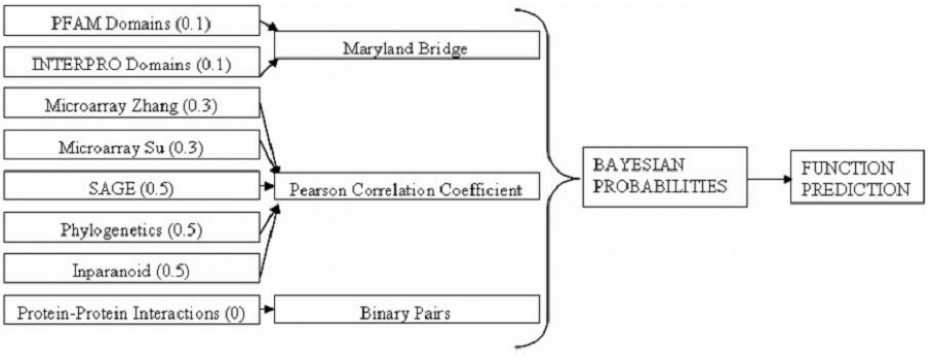


Figure 1: Flowchart of function prediction method.

Ontology	Evaluation Category	Novel Set AUC	Test Set AUC
GO_BP	3	0.64989	0.74850
	11	0.67968	0.78512
	31	0.69170	0.80624
	101	0.66368	0.75502
GO_CC	3	0.65043	0.80907
	11	0.67532	0.84761
	31	0.72454	0.83561
	101	0.64951	0.79763
GO_MF	3	0.75424	0.86666
	11	0.78520	0.89558
	31	0.81353	0.86472
	101	0.80386	0.86614

Table 1: GeneFAS performance improvement with revised thresholds for similarity measure.

3 Conclusions

GeneFAS allows users to combine the public data with their own private data and integrate both in functional inference. It is an automated tool that can provide robust and useful hypotheses for gene functions. It is freely available for download at `http://digbio.missouri.edu/genefas`.

References

[1] Barutcuoglu Z, et al. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 2006, 22(7):830–836.

[2] Chen Y and Xu D. Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*. *Nucl Acids Res*, 2004, 32:6414–6424.

[3] Deng M, et al. An integrated probabilistic model for functional prediction of proteins. *J Comput Biol*, 2004, 11:463–475.

[4] Joshi T, et al. Genome-scale gene function prediction using multiple sources of high-throughput data in yeast *Saccharomyces cerevisiae*. *OMICS: A Journal of Integrative Biology*, 2004, 8:322–333.

[5] Lanckriet GRG, et al. Kernel-based data fusion and its application to protein function prediction in yeast.

[6] Pena-Castillo L, et al. A critical assessment of *M. musculus* gene function prediction using integrated genomic evidence. *Genome Biology*, in press.

[7] The Gene Ontology Consortium. *Nature Genetics*, 2000, 25:25–29.

Computational Studies of Lens Regeneration Under Influence of Vitamin A and its Metabolite

Amit Nagal,[1*] O. P. Jangir[1]

The Present study support our previous finding that vitamin A can induce and accelerate lens regeneration from dorsal iris PECs not only in amphibians but also in young and adult swiss albino mice, guinea pig, rabbit and pigs. Mitogenic and dedifferentiative activity of vitamin A can be considered as key factor for lens regeneration as shown by several workers that impairing of functions of retinoid receptor inhibits lens regeneration. Main purpose of this study is to know how retinoids and its derivatives acting and interact of Retinoic acid receptor alpha and thus helping in lens regeneration. We have used bioinformatics and drug designing tools: Autodock3, cerius2, Insight II. Analysis of file gave the number of conformer generated and their respective docked energies. On the basis of best ranking conformers in term of docking energy result shows the interaction between Rxr alpha receptor with vitamin A and 9 cis retinoic acid. In the present study it can be concluded that vitamin A shows effect on dedifferentiation, proliferation and differentiation similar to 9 cis retinoic acid. and this study proves that Vitamin A acts on Retinoid X alpha receptor and it enhance lens regeneration in mammals.

[1]Developmental Bio lab, Department of Zoology, Dungar College Bikaner 334001, India.
* Corresponding author: amitkumarnagal@gmail.com

Structure-Based Approach for Predicting Kinase Substrates: Role of Solvent Accessibility of the Site of Phosphorylation

Narendra Kumar,[1] Debasisa Mohanty[2]

1 Introduction

Protein-protein interactions are important for almost every cellular process including replication, transcription, translation, signal transduction, immune responses and cell growth. Often such interactions involve recognition of short contiguous peptide stretch within one protein by the other interacting partner. Protein kinases and MHC are two such important classes of proteins which recognize their substrates as short peptides. Hence, theoretical methods for predicting substrates for kinases or MHCs attempt to identify substrate peptides for these proteins. Although a number of sequence motif based computational approaches are available currently, most of them are trained on the experimentally available peptide binding data, and hence can make predictions only for those classes for which substantial amount of experimental data is available.

On the basis of analysis of available protein-peptide complexes of kinases and MHCs, we have developed MODPROPEP (http://www.nii.res.in/modpropep.html) [4], a software for the modeling, visualization, and detailed atomic level analysis of protein-peptide complexes involving kinases and MHC molecules. MODPROPEP also predicts the substrates for any given protein kinase and MHC protein using residue-residue statistical pair potential as the scoring function for ranking the protein-peptide complexes. For 10 major kinase families, MODPROPEP could rank the actual binding peptide within the top 30% of all the other potential binding peptides, in more than 60% of cases. Similarly, for 90 class I MHC-peptide complexes, the true binder peptides was predicted among top 30% in 61% of cases. Unlike sequence based methods, MODPROPEP does not use any experimental data for training and it is entirely based on structural properties of protein-peptide recognition. However, MODPROPEP performs significantly better or similar to the other sequence based methods which are trained on experimental data. Therefore, MODPROPEP can potentially be used to predict substrates for newly identified kinases or new MHC alleles.

For further improving the prediction accuracy, we investigated the factors others than structural and chemical complementarity of protein and peptides involved in interaction. In case of protein kinases, surface accessibility of the substrate peptides is one such criterion. We currently rank all possible peptides with potential phosphorylation sites from a putative protein. However some of these sites might be buried deep inside protein interior, and hence not available for the phosphorylation by kinase in absence of major conformational change. Although some computational method for prediction of phosphorylation sites take into account solvent accessibility of peptides, the importance of solvent accessibility has not been analyzed thoroughly [2, 5]. To investigate the importance of the surface accessibility in the phosphorylation event by protein kinase, we systematically analyzed the solvent accessibilities of phosphorylation sites in known substrate proteins, and compared with the accessibilities of sites which are not phosphorylated.

2 Methods

The protein sequences of experimentally verified phosphorylation sites were gathered from Phospho.ELM database [1]. Since no information about the availability of crystal structures of these proteins was available, we identified their structural homologes by carrying out the blast alignment against protein sequence entries in PDB. Proteins with significant alignment over reasonable length were chosen for further analysis. We further refined the dataset by choosing only those proteins which showed conservation of the phosphorylation site in the protein structure. The protein structures corresponding to these alignments were chosen for the calculation of the solvent accessibilities of the phosphorylation sites and

[1]National Institute of Immunology, Aruna Asaf Ali Road, New Delhi, India. Email: narendra@nii.res.in
[2]National Institute of Immunology, Aruna Asaf Ali Road, New Delhi, India. Email: deb@nii.res.in

the all other serine, threonine and tyrosine containing 7 mer peptides. The solvent accessible area was calculated using NACCESS program [3]. The solvent accessible area of actual phosphorylation sites and all other ser/thr/tyr containing peptides were compared.

3 Results

A comparison of solvent accessible surface areas of serine, threonine and tyrosine residue containing stretches in the structural homologes of experimentally verified substrate protein showed that the phosphorylation sites are significantly more exposed compared to serine, threonine and tyrosine containing 7 mer stretches which are not phosphorylated. This trend was more prominent in the case of serine containing peptides (Figure 1). In view of these results, incorporation of solvent accessibility term along with the current scoring function based on the residue-residue statistical energy has the potential for further improvement in the prediction accuracy of prediction program. Our analysis also showed few interesting cases where a ser/thr/tyr containing sites with very low accessibility were also phosphorylated. It will be interesting to analyze these proteins for determining the importance of the conformational flexibility associated with phosphorylation.

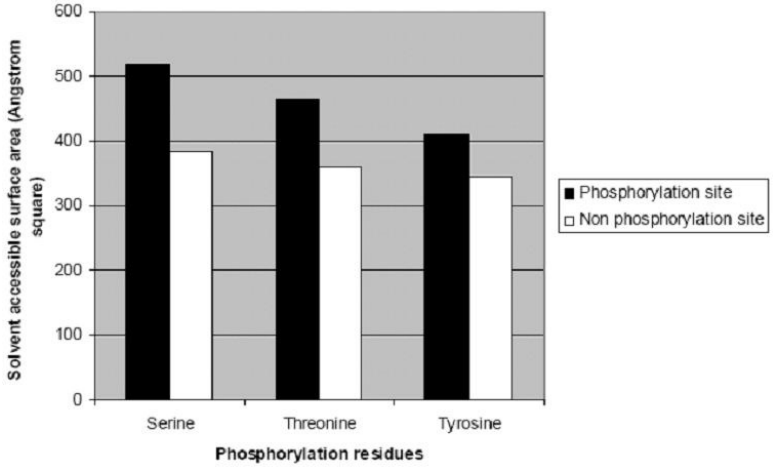


Figure 1: Average solvent accessible surface areas of serine, threonine and tyrosine residues as calculated by NACCESS program.

References

[1] Diella, F., Cameron, S., Gemund, C., Linding, R., Via, A., Kuster, B., Sicheritz-Ponten, T., Blom, N. and Gibson, T. J. 2004. Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, 5:79.

[2] Gnad, G., Ren, S., Cox, J., Olsen, J. V., Macek, B., Oroshi, M. and Mann, M. 2007. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol*, 8(11):R250.

[3] Hubbard,S. and Thornton,J. M. 1993. NACCESS Computer Program. Department of Biochemistry and Molecular Biology, University College, London.

[4] Kumar, N. and Mohanty, D. 2007. MODPROPEP: A program for knowledge-based modeling of protein-peptide complexes. *Nucleic Acids Res*, 35:W549–W555.

[5] Obenauer, J. C., Cantley, L. C. and Yaffe, M. B. 2003. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*, 31(13):3635–3641.

In Silico Modeling of Pesticidal Crystal-Like Protein Cry16Aa from *Clostridium bifermentans*

Jayasree Ganugapati,[1] Ravindra Babu Potti, Ashok Chakravarthy

1 Introduction

The family of genes coding for Pesticidal crystal-like protein is the Cry gene family. Cry genes of Bacillus thuringiensis are found in Clostridial sps. These genes are present in 80% of the Clostridium bifermentans strains tested [3]. A common characteristic of the cry genes is their expression during the stationary phase. Their products generally accumulate in the mother cell compartment to form a crystal inclusion that can account for 20–30% of the dry weight of the sporulated cells. When the inclusions are ingested by insect larvae, the alkaline pH solubilizes the crystal. The protoxin is then converted in to an active toxin after processing by the host proteases present in the midgut and causes the cell swelling, lysis and eventually death of the insect [9].

2 3D Model Building

The initial model of cry16Aa was built by using homology-modeling methods and the MODELLER software, a program for comparative protein structure modeling optimally satisfying spatial constraints derived from the alignment and expressed as probability density functions (pdfs) for the features constrained. The pdfs restrain C^{α}–C^{α} distances, main-chain N–O distances, main-chain and side-chain dihedral angles. The 3D model of a protein is obtained by optimization of the molecular pdf such that the model violates the input restraints as little as possible. The molecular pdf is derived as a combination of pdfs restraining individual spatial features of the whole molecule. The optimization procedure is a variable target function method that applies the conjugate gradients algorithm to positions of all non hydrogen atoms [8]. The query sequence from Clostridium bifermentans was searched to find out the related protein structure to be used as a template by the BLAST (Basic Local Alignment Search Tool) [1] program against PDB (Protein Databank), Table 1. Sequences that showed maximum identity with high score and less E-value were aligned and were used as a reference structure to build a 3D model for cry16Aa. The coordinates for the structurally conserved regions (SCRs) for cry16Aa were assigned from the template using multiple sequence alignment, based on the Needleman-Wunsch algorithm [7]. The structure having the least modeler objective function obtained from the modeler was improved by energy minimization; Fig. 2. The final structure obtained was analyzed by Ramachandran's map (Fig. 1) using PROCHECK (a program to check the stereochemical quality of protein structures) [6]. A comparative assessment of secondary structure was obtained using GOR IV, HNN and SOPMA.It revealed greater percentage of residues as alpha helix and random coils against the beta sheets [2, 4, 5]; Table 2.

3 Results

% of residue in most favored regions	86.1
% of residue in the additionally allowed zones	10.6
% of residue in the generously regions	2.6
% of residue in disallowed regions	0.7

Figure 1: Ramachandran plot calculations on 3D model of cry16Aa computed with the PROCHECK program.

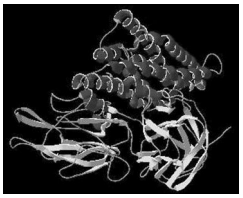


Figure 2: Energy minimised structure of crystal protein cry16Aa from Clostridium.

[1]Department of Biotechnology, Sreenidhi Institute of Science and Technology (SNIST), Yamnampet, Ghatkesar, Hyderabad 501 301, Andhra Pradesh, India. Email: jayasree1097@rediffmail.com

PDB	Protein	Chain	Identity to cry16Aa(%)
1DLC	Crystal structure of insecticidal delta-endotoxin from 2 bacillus thuringiensis at 2.5 angstroms resolution	A	24%
1JI6A	Crystal structure of the insecticidal bacterial delta 2 endotoxin cry3bb1 bacillus thuringiensis	A	23%
1CIY	Insecticidal toxin: structure and channel formation	A	21%

Table 1: Data for closest homologue for cry16Aa with known 3D structure obtained with the blast server against PDB.

Prediction Tool	Alpha Helix	3_{10} Helix	Beta Bridge	Extended Strand	Beta Turn	Random Coil	Ambigous States	Other States
GOR4	17.78%	0.00%	0.00%	26.26%	0.00%	55.95%	0.00%	0.00%
HNN	32.63%	0.00%	0.00%	17.13%	0.00%	50.24%	0.00%	0.00%
SOPM	29.53%	0.00%	0.00%	25.29%	7.83%	37.36%	0.00%	0.00%

Table 2: Secondary structure prediction, percentage of helices, sheets and random coils.

References

[1] Altschul SF, Madden TL, Schaffer AA, et al. 1997. Gapped BLAST and PSIBLAST: A new generation of protein database search programs. *Nucleic Acids Res*, 25, 3389–3402.

[2] Combet C, Blanchet C,Geourjon C, and Deléage. 2000. G.NPS@:NetworkProteinSequenceAnalysis. 25, 147–150.

[3] Frédérique Barloy, Marguerite M. Lecadet, Armelle Delécluse. 1998. *Current Microbiology*, 36.

[4] J. Garnier, J.-F. Gibrat, B. Robson. 1996. GOR secondary structure prediction method version IV. *Methods in Enzymology*, 266, 540–553.

[5] Geourjon C and Deléage G. 1994. SOPM: A self-optimised method for protein secondary structure prediction. *Protein Engineering*, 7, 157-16.

[6] Laskoswki RA, MacArthur MW, Moss DS and Thornton JM. 1993. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 26, 283–291.

[7] Needleman SB and Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48, 443–453.

[8] Sali A and Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, 234, 779–815.

[9] Schnepf HE, et al. 1998. Bacillus thuringiensis and its pesticidal crystal proteins. *Microbiology and Molecular Biology Reviews*, 62, 775–806.

Patterns of Differential Over-Expression of the Oncogene IKBKE in HER2+ and Basal Breast Cancer

Gabriela Alexe,[1] Erhan Bilal,[2] Nilay Sethi,[3] Lyndsay Harris,[4] Vasisht R. Tadigotla,[5] Shridar Ganesan,[6] Gyan Bhanot[7]

1 Introduction

This paper presents new developments on the role of the NF-κB pathway in subtypes of breast cancer following from a recent paper by Boehm et al [2]. In [2] it was shown that IKBKE, a kinase in the NF-κB pathway, is a breast cancer oncogene and is over-expressed in a subset of breast cancers. Using microarray data from Wang et al [6], we have recently shown [1] that node negative breast cancers treated with surgery and radiation but no adjuvant or neo-adjuvant therapy separate into at least eight distinct subtypes characterized by differential expression of genes and different rates of long term metastasis free survival. In particular, HER2+ breast cancers split into two subtypes, one of which (HER2+$_I$), has a significantly low long term recurrence rate (11% vs 48%) compared to the other subtype HER2+$_{NI}$ correlated with an overexpression of immunoglobulins, cytokine and chemokine genes of the adaptive immune system. This suggests that for HER2+ breast cancers, the presence of a lymphocytic infiltrate in the tumor environment correlates with improved natural history. Using microarray and paraffin sections from a neo-adjuvant HER2+ trial [3], we have verified this correlation between our HER2+I subtype and a lymphocytic infiltrate using a blind study involving pathologists at two different institutions.

We also find that Basal-like (triple negative) tumors separate into two subtypes (BA1 and BA2). These subtypes correlate well with the subtypes found in a recent study [5] that showed that basal-like tumors separate into a set with X isodisomy and another with Xp isodisomy. Our clustering analysis also finds that the BA1 subtype is characterized by up-regulation of IFN genes in the innate immune system, suggesting that it also elicits a differential immune response compared to the BA2 subtype.

In the present paper, we show that in the IKBKE oncogene is upregulated only in HER2+$_I$ and Basal-like breast cancer subtypes, suggesting that the immune signature seen in these subtypes may be linked to the NF-κB pathway.

2 Results

Our observations on the Wang et al dataset suggest a connection between the NF-κB pathway and immune system activation in some of the subtypes identified in our clustering analysis. To see if this can be validated, we re-examined the microarray data of [4] to determine the expression level of IKBKE among the 8 subtypes of breast cancer. We found that IKBKE is up-regulated in the Basal-like and HER2+$_I$ subtypes only. These results are summarized in Table 1. The other genes shown are those that are highly correlated with IKBKE expression in the various subtypes.

3 Materials and Methods

The method we have developed [1] to analyze microarray data and identify subtypes of disease first uses several techniques such as signal-to-noise ratio, t-test and principal component analysis to reduce the set of genes to a subset which are useful in stratifying the data into subtypes. Next, using this subset of genes, we use statistical measures to identify the optimum number of clusters in the data and

[1]The Broad Institute of MIT and Harvard, Cambridge, MA, USA. Email: galexe@broad.mit.edu
[2]Rutgers University, NJ, USA. Email: ebilal@rutgers.edu
[3]Robert Wood Johnson Medical School and UMDNJ, New Brunswick, NJ, USA. Email: nsethi@princeton.edu
[4]Yale Cancer Center, Yale University, New Haven, CT, USA. Email: lyndsay.harris@yale.edu
[5]Department of Physics, Boston University, Boston, MA, USA. Email: vasisht@buphy.bu.edu
[6]Cancer Institute of New Jersey, New Brunswick, NJ, USA. Email: ganesash@umdnj.edu
[7]Institute for Advanced Study, Princeton, NJ 08540, USA. Email: gyanbhanot@gmail.com

separate the samples into these subtypes using consensus ensemble clustering. This approach averages over many clustering methods and data perturbations to produce an agreement matrix which can be sorted to separate the samples into clusters that are robust to perturbations of sampling bias and gene and clustering method choice.

4 Discussion and Future Prospects

These results suggest that the IKBKE oncogene disregulates the NF-κB pathway only in two subtypes of breast cancer (basal-like and HER2+$_I$). To validate this claim by in-vivo experiments on tissue microarray of samples classified into clinical subtypes of HER2+, basal-like and luminals, we have established a reliable positive control for NF-κB activation by comparing immunostaining for NF-κB in two breast cancer cell lines that have increased IKBKE activity (MCF-7, MDA-MB-453) compared to a cell line (MCF-10A) that has a normal expression of IKBKE as observed in the study by Boehm et al [2]. Additionally, we isolated protein from these cell lines and immunoblotted for IKBKE to support our immunofluorescence data. Paraffin embedded sections for the cell lines MDA-MB-453 and MCF-10A were created and tested as positive and negative controls, respectively, for IKBKE and NF-κB activation. These experiments were successful and resulted in positive and negative controls for IKBKE which could be used as benchmarks for tissue microarray samples from patients.

We are currently studying tissue microarray from ∼300 patient tumor samples with known ER, PR, HER2 status using IRB approved protocols. We will use paraffin-embedded MDA-MB-453 as positive control and paraffin-embedded MCF-10A as negative control to study whether the hypothesis that only Basal-like and HER2+$_I$ samples show an up-regulation of IKBKE and activation of the NF-κB pathway. Future directions include in-vivo experiments that will analyze the role of IKBKE in basal-like breast cancer by inoculating mice with a basal-like breast cancer cell line compromised for IKBKE activity.

References

[1] Alexe G, Dalgin GS, Scanfeld D, et al. 2007. High expression of lymphocyte-associated genes in node-negative HER2+ breast cancers correlates with lower recurrence rates. *Cancer Res*, 67(22):10669–10676.

[2] Boehm JS, Zhao JJ, Yao J, et al. 2007. Integrative genomic approaches identify IKBKE as a breast cancer oncogene. *Cell*, 29(6):1065–1079.

[3] Harris LN, You F, Schnitt SJ, et al. 2007. Predictors of resistance to preoperative trastuzumab and vinorelbine for HER2-positive early breast cancer. *Clin Cancer Res*, 13(4):1198–1207.

[4] Ivshina AV, George J, Senko O, et al. 2006. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res*, 66(21):10292–10301.

[5] Richardson AL, Wang ZC, De Nicolo A, et al. 2006. X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell*, 9(2):121–132.

[6] Wang Y, Klijn JG, Zhang Y, et al. 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 365(9460):671–679.

	ALL	CA	BA	BA1	BA2	HER2	HER2I	HER2NI	Lum	LumA	LumB	N
IKBKE	0	0	0.6	0.67	0.5	-0.19	0.57	-0.57	-0.13	-0.48	-0.07	-0.8
CCL5	0	0	0.7	1.23	0.2	0.23	0.42	-0.35	-0.32	-0.15	-0.39	-0.3
E2F3	0	0	1.2	0.99	0.38	-0.11	0.08	-0.24	-0.35	-0.56	-0.31	-1.4
IRF1	0	0	0.4	1.09	-0.2	0.18	1.34	-0.57	-0.2	-0.16	-0.25	-0.2
RELB	0	0	0.7	1.09	0.32	0.04	0.67	-0.24	-0.27	-0.07	-0.32	-0.1

Table 1: Relative expressions of IKBKE and correlated genes in subtypes. Upregulated gene expression values are highlighted red. The subtypes are as follows: BA = basal-like; BA1 = basal-like with innate immune signature; BA2 = basal-like w/o innate immune signature; HER2I = HER2+ with lymphocyte infiltrate signature; HER2NI = HER2+ w/o lymphocyte infiltrate signature; Lum = luminals (ER+, PR+, HER2-); LumA = luminal A; LumB = luminal B.

Parameter Estimation of Oscillatory Systems

Kok Siong Ang, Rudiyanto Gunawan[1]

1 Introduction

Oscillatory systems are found in various processes of biological systems. Examples of such oscillatory behaviour include the circadian rhythm (PER and TIM proteins) and the p53-mdm2 oscillations during DNA damage. This work discusses the application of global optimization methodology for parameter estimation of such oscillatory systems. A novel cost function is developed to facilitate the parameter search and applied to a circadian rhythm model[3].

2 Theory

Oscillatory biochemical systems form an important area within Systems Biology. Such systems are typically modeled as limit cycles with coupled ordinary differential equations (ODEs):

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}(t),\mathbf{p}) \tag{1}$$

An oscillatory system has 2 characteristics: shape and period, both of which can be the source of mismatch between the model and data. In contrast, non-oscillatory systems have only the first characteristic of shape. The characteristic of period gives rise to additional difficulty to the typical parameter estimation problem. This can be illustrated by inspecting the parametric sensitivity of the state x_i to parameter p_j [4]:

$$\frac{\partial x_i}{\partial p_j} = \left(\frac{\partial \mathbf{x}}{\partial p_j}\right)_{\tau} - \frac{t}{\tau}\frac{\partial \tau}{\partial p_j}\frac{dx_i}{dt} \tag{2}$$

where τ is the system period and t is time. The first term contains the parametric state sensitivity with respect to constant period, and the second term contains $(\partial\tau/\partial p_j)$, the parametric period sensitivity. However, the more interesting feature is the presence of (t/τ) within the second term. It is thus obvious that as $t \to \infty$, the second term will blow up as well. Thus this gives rise to problems when attempting parameter estimation with (time-based) data.

3 Parameter Estimation

3.1 Cost Function

In parameter estimation, the most common approach is to construct the cost function by summing the square of the errors between the data and predicted values, or the least squares method. As mentioned, the error for an oscillatory system contains error due to the shape and error due to period. We sought to separate them by converting time based data into phase based data. Dividing time by the system period (t/τ) accomplishes this.

The cost function is now formed with the 2 errors:

$$\min_{\mathbf{p}} \left\{ \omega_1 \sum_{k=1}^{Nx} \sum_{j=1}^{Ns_k} \sum_{i=1}^{Nr_{k,j}} \left(\hat{x}_{ijk} - x(\mathbf{p})_{ijk}\right)^2 + \omega_2 \sum_{h=1}^{Nx} \left(\hat{\tau}_h - \tau_h(\mathbf{p})\right)^2 \right\}. \tag{3}$$

where ω_1 and ω_2 are weights for the error in data points and period respectively. The choice of weights can be arbitrary but suitable values can be the number of data points per period. The errors are summed over the number of readings at each time point for each particular state $Nr_{(k,j)}$, the number time points at during which measurements are made for each state $Ns_{(k)}$, and the number of states Nx.

[1]Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore. Email: chegr@nus.edu.sg

3.2 Search Algorithm

As in [1], the parameter estimation problem, with the cost function developed above, is states as a nonlinear programming problem and a global optimization algorithm is applied. For this work the Differential Evolution algorithm [2] was used.

Though [1] showed that the Differential Evolution algorithm was inefficient for its case study, satisfactory performance was obtained for the case study within this work.

4 Results and Discussion

To test the efficacy of the method, it is applied on a 2 state, 9 parameter circadian model [3]. The data is generated *in-silico* with the same model with noise added. Since the model parameters are known, it is easy to verify success or failure.

Table 4 shows the parameters and the corresponding scores and periods obtained via parameter estimation. Figure 1 shows the fit between the data and the estimated system (Run 1). It can be seen that an excellent fit can be obtained even if not all the parameters match the original. Further investigation by identifiability analysis will help in explaining the large discrepancies between the original and estimated of certain parameters.

	ν_m	k_m	ν_p	k_{p1}	k_{p2}	k_{p3}	K_{eq}	P_{crit}	J_p	Score	Period
Actual	1.0	0.1	0.5	10	0.03	0.1	200	0.1	0.05	-	24.67146
Run 1	1.03633	0.10684	0.46525	18.438	0.093544	0.051336	632.67	0.12955	0.084682	0.002888	24.29158
Run 2	1.01958	0.10140	0.48119	22.494	0.088638	0.057906	903.69	0.10909	0.077047	0.002903	24.27552
Run 3	1.01504	0.10454	0.47588	21.739	0.098599	0.050996	825.17	0.12227	0.083191	0.002892	24.23535
Run 4	0.99868	0.10553	0.47245	16.805	0.072737	0.054970	459.61	0.12836	0.088153	0.002892	24.28700

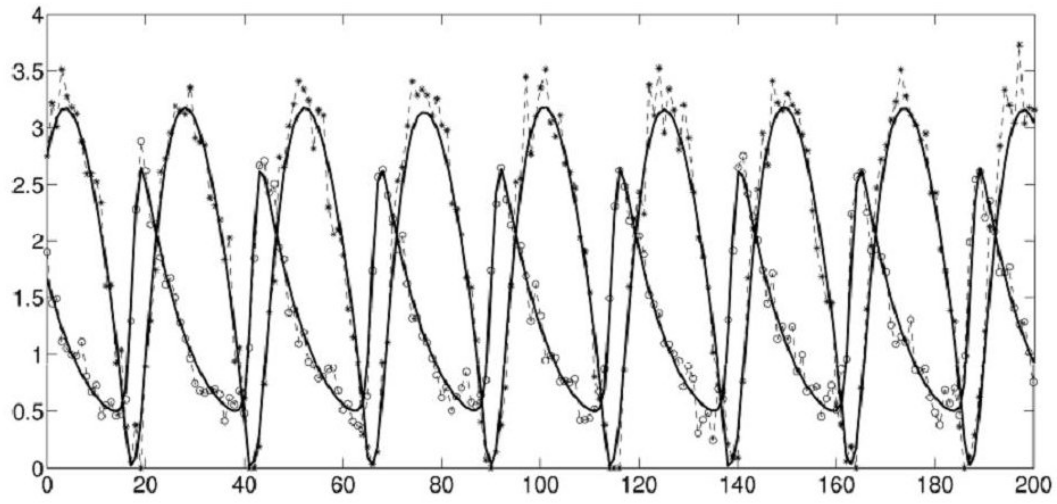


Figure 1: Predicted (solid lines) and data set (dash lines) for the system

References

[1] Moles, C. G., Mendes, P., and Banga, J. R. 2003. Parameter estimation in Biochemical Pathways: A Comparison of Global Optimization Methods. emphGenome Research 13:2467-2474.

[2] Price, K. V., Storn, R. M. and Lampinen, J. A. 2005. *Differential Evolution - A Practical Approach to Global Optimization*. Berlin: Springer.

[3] Tyson, J. J., C. I. Hong, C. D., Thron and B. Novak. 1999. Simple model of circadian rhythms based on dimerization and proteolysis of PER and TIM. *J. Biol. Rhythms.* 13:70-87.

[4] Zak, D. E., Stelling, J. and Doyle, F. J. 2005. Sensitvity analysis of oscillatory (bio)chemical systems. *Computers & Chemical Engineering* 29:663-673.

Aging Studies: A Stochastic Approach

Suresh K Poovathingal,[1] Rudiyanto Gunawan,[1]
Jan Gruber,[2] Barry Halliwell[2]

1 Introduction

Mitochondria are the powerhouses of eukaryotes, but at the same time they also produce mutagenic reactive oxygen species (ROS) as the byproducts of cellular respiration. The accumulation of mitochondrial DNA (mtDNA) mutations has been postulated to result from the proximity of the mtDNA to the mitochondrial electron transfer chain (ETC) which is thought to be the major source of ROS. The loss of mitochondrial function due to such mutations has been associated with metabolic and degenerative diseases, whose clinical symptoms progress with ageing [3]. The connection between mtDNA mutations and ageing has been supported by evidence showing that the amount of somatic mtDNA mutations increased with age [1]. Furthermore, mice with defective mtDNA polymerase show the symptoms resembling premature aging [2]. In this work, the consequence of inherent biological stochasticity on the maintenance of genomic stability of mtDNA is investigated through the development of an in silico model.

Numerous mathematical models have been proposed to explain the accumulation of somatic mutations in mtDNAs. Kowald and Kirkwood have done several pioneering works in the field of ageing modeling [3]. One important assumption made by previous investigators is the existence of a ROS "vicious cycle" theory; ROS causes mtDNA mutations and in turn these mutations cause higher production of ROS. This hypothesis however has been a source of intense debate over decades, due to the lack of experimental evidence for the existence of such a vicious cycle mechanism. Today it is clear that many of the original assumptions that led to the vicious cycle model are unfounded [9]. Furthermore, most of the existing models were deterministic (ODEs). A few exceptions include Langevin-type stochastic models due to Samuels and Chinnery [4], which were based on the assumption of relaxed replication of mtDNA from cell cycle. Here, mitochondrial fusion and fission were assumed to occur frequently enough to justify a single well-mixed mtDNA pool in the cell. This assumption is also used in the present model. One weakness of the existing models is the large number of unknown parameters that need to be set or estimated. In the work presented here, we developed a minimal chemical master equation model of mtDNA somatic mutation which can capture features of experimental data on mouse [2].

2 Model Description

The model captures two main processes that contribute to the maintenance of mtDNA genomic stability, namely the degradation and replication processes:

$$mtDNA \rightarrow \emptyset \quad a = \quad k_d \cdot mtDNA \tag{4}$$

$$mtDNA \rightarrow 2mtDNA \quad a = \quad v_r \left(1 - \frac{(W+M)^n}{K^n + (W+M)^n}\right) \tag{5}$$

where k_d, v_r, K, and n are the model parameters. The a's are known as the propensity functions, for which $a \times dt$ gives the probability that a given process takes place in the time range $[t, t+dt)$. The model further tracks the number of wildtype (W) and mutant (M) mtDNAs individually. In addition, there exist a probability (k_m) that a replication of W will produce one W and one M, depicting a mutation process during the replication process. As the mtDNA replication is known to be regulated by nucleus, perhaps in response to the energetic needs of the cell, the propensity function for mtDNA replication depends on the combination of $(W + M)$ (non pathogenic mutations) according to a Hill-type function.

In order to deal with the stochastic elements inherent in cellular processes [7], Stochastic Simulation Algorithm (SSA) was used to simulate the model [6]. A modified version of this algorithm has been used for the present work, which is not detailed here for brevity. The parameters were obtained from reported values based on experimental data for mouse, in which $k_d = 2.3377 \times 10^{-3} day^{-1}$ [5], $k_m =$

[1]Department of Chemical & Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117576. Email: suresh.poovathingal@nus.edu.sg, chegr@nus.edu.sg

[2]Department of Biochemistry, Neurobiology and Ageing Programme, Centre for Life Sciences (CeLS), 28 Medical Drive, Singapore 117456. Email: jan_gruber@nus.edu.sg, bchbh@nus.edu.sg

1.8×10^{-6} replication^{-1} [7] and $W_0 = 3000$ (Brain cells); $W_0 = 3500$ (Heart) [8]. The remaining parameter were either dependent on the above parameters or were specific for the present model. The stochastic simulations realization were done in IBM Blade Linux computing cluster with 112 processors.

3 Results and Discussion

A comparison of the frequency of mutations as predicted by the model and a recent experiment in mouse is illustrated in Figure 1a. In this experiment, mutations on the *TaqI* restriction site (TCGA) located in the gene encoding region of the 12S rRNA (bp 634-637) subunit was tracked over 35 months [2]. The simulation and the experimental results over 36 months were in good agreement. Although a best fit analysis of both simulations and experiments using the same small sample numbers suggested that the mutational burden accumulated exponentially with time, careful analyis of the model further revealed that the average of the mutation frequency actually followed a linear dependence; see dotted line, Figure 1b. According to Figure 1b, both the mean and median of the mutation frequency follow an linear dependence with age. The apparent exponential increase in the mutational burden was an artifact of a small number of data taken from long-tailed distributions.

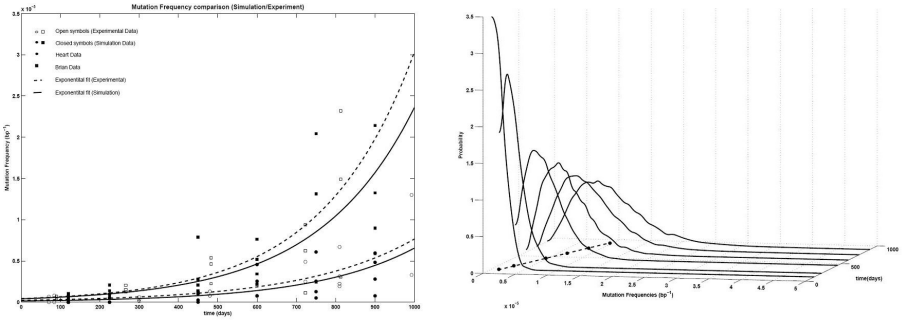


Figure 1: (a) Figure illustrating the exponential fit for the Mutation frequencies (Simulation/Experimental) for different age length. (b) Evolution of histogram of the mutation frequencies for different age lengths.

4 Conclusions

Precisely how a cell maintains its mtDNA population is fundamentally important to our understanding the relationship between the mitochondrial genome, ageing, and diseases. In this work, a minimal model, accounting only degradation, replication, and mutation of mtDNAs, is developed and shown to have good agreement with experiments. Despite the apparent exponential increase of mtDNA mutation load with time, the model predicts that the average mutation frequency to grow linearly with age. Importantly, we were able to reproduce recent experimental results without the assumptions of either a vicious cycle or a proliferative advantage for mutant mtDNA. The model provides a starting point for the development of more complex and realistic representations of mtDNA somatic mutation, including the role of mitochondrial fusion and fission.

References

[1] Khaidakov, M., Heflich, M. H., Manjanatha, M. G., Myers, M. B., and Aidoo, A. 2003. Accumulation of point mutations in mitochondrial DNA of aging mice. *Mutat. Res.*, 526(1–2):1–7.

[2] Vermulst, M., Bielas, .J .H., Kujoth, G. C., Ladiges, W. C., Rabinovitch, P. S., Prolla, T. A., Loeb, A. 2007. Mitochondrial point mutations do not limit the natural lifespan of mice. *Nat. Genet.*, 39(4):540–543.

[3] Kowald, A. and wood, T. B. 1994. Towards a network theory of ageing: a model combining the free radical theory and the protein error theory. *J. Theor. Biol.*, 168(1):75–94.

[4] Elson J.L., Samuels, D.C., Turnbull D.M., and Chinnery P.F. 2001. Random intracellular drift explians the clonal expansion of mitochondrial DNA mutations with age. *Am. J. Hum. Genet.*, 68:802–806.

[5] Collins, M. L., Shannon, E., Hoh, R., Hellerstein, M. K. 2003. Measurement of mitochondrial DNA synthesis in vivo using a stable isotope-mass spectrometric technique. *J. Appl. Physiol.*, 94:2203–2211.

[6] Gillespie D. T. 1977. Exact Stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361.

[7] Coller, H. A., Khrapko, K.,Bodyak, N. D., Nekhaeva, E., Herrero-Jimenez, P., Thilly, W.G. 2001, High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. *Nat. Genet.*, 28:147–150.

[8] Wiesner, R. J., Reuegg, J. C., Morano, I. 1992. Counting target molecules by exponential polymerase chain reaction: Copy number of mitochondrial DNA in rat tissues. *Biochem. Bioph. Res. Co.*, 183(2):553–559.

[9] Wiesner, R. J., Zsurka, G., Kunz, W. S. 2006. Mitochondiral DNA damage and aging process-facts and imaginations. *Free Radical Res.*, 40(12):1284–1294.

GlycoVault: An Online Storage and Visualization System for Glycan Structures

Faraaz N. K. Yusufi,[1] Satty Ganeswara Reddy,[1] May May Lee,[1] Dong-Yup Lee[1,2]

1 Introduction

Glycans are complex chains of monosaccharides that play critical roles in several structural and modulatory functions in cells. Although glycans are considered one of the most important classes of molecules after DNA and proteins, the development of informatics methods to support and advance their research has lagged behind those available for other types of data. It is only in recent years that there has been an increase in the availability of informatics resources such as glycan databases and algorithms for analyzing glycan structures and their interactions [1]. Despite the deficiency in informatics tools, high throughput technologies have only led to the production of ever increasing amounts of glycan data. Glycobiology labs are currently using several different technologies to produce many different types of data. Unfortunately this diversity in data makes it difficult to create a central storage system and information is saved as a jumble of spreadsheets and text files. Further compounding the problem is the visual nature of glycan data, with many labs resorting to storing structure information as hand-drawn annotation on printouts of spreadsheets. In order to address the need for a centralized storage and visualization platform for glycan data we developed the web-based GlycoVault system.

2 Features

GlycoVault allows users to upload and store experimental glycan data such as GlycoMod [2] files along with some annotation data. These files can be later retrieved and used to generate visual reports listing figures of glycan structures observed in an experiment. GlycoVault also contains several interactive tools to visually explore glycan structures. A drawing tool is available to interactively draw and store glycan structures. The glycosylation reaction network can be thought of as a graph with the nodes representing glycan structures and edges showing possible enzymatic reactions. GlycoVault allows users to create networks of glycan structures and then visualize pathways connecting different structures. Users can also determine what reactions are necessary to convert one glycan structure into another.

3 Implementation

GlycoVault uses the GlycoDigit format to represent glycan structures internally. GlycoDigit is a fixedlength alpha-numeric code for representing glycan structures commonly found in secreted glycoproteins. The code uses a pre-assigned alpha-numeric index to represent the monosaccharides attached in different branches to the core glycan structure. The numeric nature of the code makes it ideal for the development of a mathematical operators and algorithms to compare glycan structures. GlycoVault is implemented through Java Server Pages (JSP) and uses the MySQL database to store information. The interactive visualization tools are developed using Adobe Flex.

References

[1] Pérez, S. and Mulloy, B. 2005. Prospects for glycoinformatics. *Curr. Opin. Struct. Biol.*, 15:517–524.

[2] Cooper C.A., Gasteiger E. and Packer N. 2001. GlycoMod: A software Tool for Determining Glycosylation Compositions from Mass Spectrometric Data. *Proteomics*, 1:340–349.

[1]Bioprocessing Technology Institute, Biomedical Sciences Institute, 20 Biopolis Way, #06-01 Centros, Singapore 138668.
[2]Department of Chemical and Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117576. Email: cheld@nus.edu.sg

MFAML: Metabolic Flux Analysis Markup Language

Jong Myoung Park,[1,3] Hongseok Yun,[1,2,3] Sang Yup Lee[1,2,3]

Recent advances in bioinformatics have led to the need for information standards to define, share, and evaluate computational models of complex biological systems. For this purpose, international communities and research teams have been developing several eXtensible Markup Language (XML)-based modeling languages. However, to our knowledge there is no standard format suitable for implementing metabolic flux analysis, which is one of the most widely adopted techniques for quantitative analysis of metabolic fluxes. This paper describes a new modeling language, *Metabolic Flux Analysis Markup Language* (MFAML), designed for the formal representation of metabolic flux models, and presents an open framework for the effective exchange of such models. It communicates basic information with System Biology Markup Language (SBML) models and provides additional data structures for MFA: balancing constraints, flux variables and objective function. MFAML also provides an Application Programming Interface (API) for converting the models to a variety of Linear Programming (LP) format. It makes it possible to use any efficient solver as modelers may want to do. With these functionalities, MFAML makes the pipeline from modeling to analyzing the metabolic networks.

Acknowledgments. This work was supported by the Korean Systems Biology Research Program M10309020000-03B5002-00000 from the Ministry of Science and Technology. Further supports by LG Chem Chair Professorship, Microsoft and IBM SUR program are appreciated.

[1]Department of Chemical and Biomolecular Engineering (BK21 Program), Metabolic and Biomolecular Engineering National Research Laboratory, Institute for the BioCentury, KAIST, Daejeon 305-701, Republic of Korea.

[2]Department of Bio and Brain Engineering, BioProcess Engineering Research Center and Bioinformatics Research Center, KAIST, Daejeon 305-701, Republic of Korea.

[3]Center for Systems and Synthetic Biotechnology, Institute for the BioCentury, KAIST, Daejeon 305-701, Republic of Korea. Email: `leesy@kaist.ac.kr`

EcoProDB: The Protein Database for *Escherichia coli*

Jong Myoung Park,[1,3] Hongseok Yun[1,2,3] Jeong Wook Lee,[1]
Joonwoo Jeong,[2] Jaesung Chung,[2] Sang Yup Lee[1,2,3]

EcoProDB database provides the information on *E. coli* proteins identified on 2-D gels along with other resources collected from various databases and published literature. The database has a comprehensive feature of showing the expression levels of *E. coli* proteins under different genetic and environmental conditions. In addition, the database has detailed information on subcelluar localization, theoretical 2-D map, and experimental 2-D map. Users can access and compare their own 2-D gels via an interactive web interface and application such as the Map Browser and the Online tools. Using EcoProDB, users can efficiently grasp the core information associated with the proteins and 2-D gel results obtained from several different experimental sets for more convenient and enhanced analysis.

Acknowledgments. This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MOST) (No. M10309020000-03B5002-00000). Further supports by LG Chem Chair Professorship, Microsoft and IBM SUR program are appreciated.

[1]Department of Chemical and Biomolecular Engineering (BK21 Program), Metabolic and Biomolecular Engineering National Research Laboratory, Institute for the BioCentury, KAIST, Daejeon 305-701, Republic of Korea.
[2]Department of Bio and Brain Engineering, BioProcess Engineering Research Center and Bioinformatics Research Center, KAIST, Daejeon 305-701, Republic of Korea.
[3]Center for Systems and Synthetic Biotechnology, Institute for the BioCentury, KAIST, Daejeon 305-701, Republic of Korea. Email: `leesy@kaist.ac.kr`

Development of an Integrative Online Tool for Modeling and Simulation of Cellular Networks

Jong Myoung Park,[1] Choamun Yun,[1,3] Hongseok Yun,[2,3] Sunwon Park,[1,3] Sang Yup Lee[2,3,4]

A web-based environment is developed for comprehensive modeling and simulation of cellular networks. WebCell provides the model library of rigorously validated and classified models that are publicly available. It also serves as the personal database for uploading and evaluation of any models of interest. The imported or created models can be validated based on thermodynamic principles and also explored with steady-state or dynamic simulations by various methods including structural pathway and metabolic control analysis. Models are allowed to be uploaded or exported in Systems Biology Markup Language (SBML) for efficient communication with other packages supporting SBML. Since its first service in 2004, WebCell has been continuously upgraded and utilized by more than one hundred registered members. The current version of WebCell 2.0 supports recently released SBML Level2 Version3 and the number of available kinetic models has been increased.

Acknowledgments. This work was supported by the Korean Systems Biology Research Program M10309020000-03B5002-00000 from the Ministry of Science and Technology. Further supports by LG Chem Chair Professorship, Microsoft and IBM SUR program are appreciated.

[1]Department of Chemical and Biomolecular Engineering (BK21 Program), Process Systems Laboratory, KAIST, Daejeon 305-701, Republic of Korea.

[2]Department of Chemical and Biomolecular Engineering (BK21 Program), Metabolic and Biomolecular Engineering National Research Laboratory, KAIST, Daejeon 305-701, Republic of Korea.

[3]Department of Bio and Brain Engineering, BioProcess Engineering Research Center and Bioinformatics Research Center, KAIST, Daejeon 305-701, Republic of Korea.

[4]Center for Systems and Synthetic Biotechnology, Institute for the BioCentury, KAIST, Daejeon 305-701, Republic of Korea. Email: `leesy@kaist.ac.kr`

Gene Expression Profiling for the Classification of Cancers of Unknown Primary

Noel G Faux,[1,*] Richard Tothill,[2,*] Justin Bedo,[1,3,*]
David Bowtell,[2] Adam Kowalczyk[1]

1 Introduction

In three to five percent of new cancer cases, the site of origin of a tumor cannot be determined by conventional methods [2] and therefore treatment and patient management is not optimal. Such cases are commonly referred to as cancer of unknown primary (CUP). Through the use of microarray gene expression profiling, we have shown that it is possible to classify primary and metastatic cancers of known origin into their respective cancer types [5]. However, the cost and time required to perform microarray experiments and efficacy of the microarrays using small and degradable amounts of sample limits the application of this technology in a clinical setting.

In addition to our work with microarrays we have also shown that it is possible to classify among 5 tumor types using less than 80 genes on a quantitative-PCR (QPCR) low-density array with support vector machine (SVM) classifiers [5]. We have now recently developed a QPCR classifier for 18 tumor types using a 786 gene set derived from analysis of three independent microarray studies [3, 4, 5]. Using a dataset representing 218 cancer samples (profiled across the 768 gene set) we evaluated several classification algorithms, including SVM and a centroid classifier, in combination with various feature selection procedures (centroid feature selection [cfs], recursive feature elimination and ttest) using leave one out, 3-fold cross-validation and bootstrap error estimation.

As the goal is a diagnostic test, suitable for clinical application it is important to develop highly accurate classifiers using relatively few features (genes). Practical limitations on the number of genes that can be assayed combined with the high number of classes and small sample size (some classes contained fewer than 10 samples) provides a challenging problem. Nevertheless we were able to induce classifiers able to predict the 18 types of primary tumors accurately using very few genes. In particular, the combination of centroid based classification and feature selection [1] achieved a good level of performance with remarkably few genes.

2 Results

Using the 218 samples as the training set we generated multi-class classifiers for each node in the tumor subtype hierarchy (Fig. 1) as well as a single multi-class classifier for all cancer types. We used a one-vs-all (OVA) strategy to generate all multi-class classifiers. Whilst we achieved reasonable accuracy for the single multi-class classifier (29% error rate), significant improvement was achieved when the classifiers were trained only within one node of the hierarchy, e.g. trained only within the epithelial (EPI) node only (17%, Table 1). Further improvement in prediction was achieved by looking at the top three predictions (relaxed error rate, Fig. 2). For all trained classifiers there was approximately a 2 fold or better improvement in accuracy (Table 1).

Node (# tumor types)	Balanced Error Rate	Relaxed Error Rate	# features (total ave unique)
All tumor types (18)	29%	14%	16 (239)
Root (2)	12%	N/A	16 (31)
Epithelial (11)	17.3%	7.7%	8 (83)
Non-Epithelial (4)	9.5%	1.9%	16 (60)
Gastric Intestinal Tract (4)	60.4%	13.9%	4 (20)

Table 1: Summary of the best performance (defined as the minimum balanced or relaxed error) for each classifier generated for the different nodes in the hierarchical tree (Fig. 1). Using centroid classification in conjunction with CFS with bootstrap 0.632+ error estimation.

[1]National ICT Australia (NICTA), Victorian Research Laboratory, The University of Melbourne, Carlton 3010, Victoria Australia.
[2]Peter MacCallum Cancer Centre, St Andrew's Place, East Melbourne 3002, Victoria. Australia.
[3]The Australian National University, ACT 0200, Australia.
* Authors contributed equally to the work.

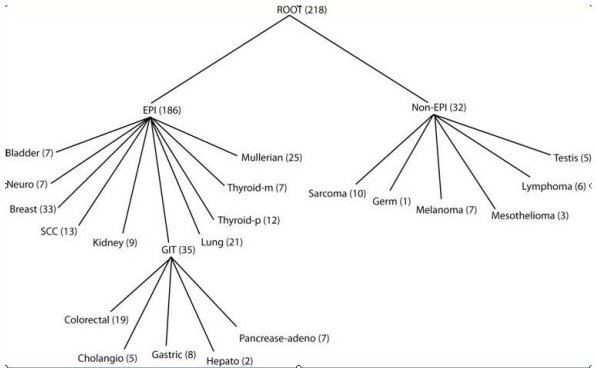


Figure 1: The hierarchical tree of the tumor types. EPI epithelial, Non-EPI non-epithelial, GIT gastrio intestinal track, SSC squamous cell carcinoma, thyroid-m medullary, thyroid-p papillary and Mullerian ovarian and uterine carcinomas. The numbers in the parenthesis are the number of samples for each cancer type. Types with 3 or less samples were excluded.

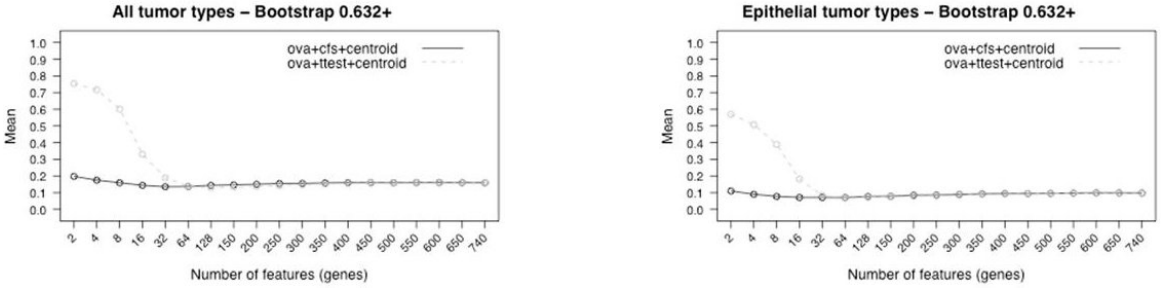


Figure 2: Line plots showing the mean relaxed error rate for the classification of all tumor types and only the epithelial subtypes. One-vs-all (ova) multi-class classification strategy was used. Using centroid classification in conjunction with cfs and ttest feature selection and performing the feature selection only for the given comparison.

By generating separate classifiers at each node there is a reduction in the total number of genes required in comparison to the single multi-class classifier, 194 genes and 239 genes respectively (Table 1). However, while there is a reduction in error rate per node (Table 1), it is unclear if a lower error rates will be realised when using a hierarchical classifier on unknown samples due to the potential problem of compound errors. Therefore further experiments are required to ascertain the hierarchical performance.

3 Summary

Here, we have shown that accurate multi-class gene expression based classifiers can be generated using a relatively small number of gene features when used in combination with quantitative real time PCR (194 genes for classification of all 18 cancer types). Increasing the number of samples available for training the classifiers (currently in progress) is expected to further increase the accuracy of the classifiers. Given the technical advantages of QPCR over microarray we believe a diagnostic test based on expression profiling and the centroid classifier is viable in a clinical setting for classifying cancers of unknown primary.

References

[1] Bedo J, Sanderson C, Kowalczyk A. 2006. An efficient alternative to SVM-based recursive feature elimination with applications in natural language processing and bioinformatics. In: *Proc. Australian Joint Conference on Artificial Intelligence*, 4304:170–180.

[2] Briasoulis E and Pavlidis N. 1997. Cancer of unknown primary origin. *Oncologist*, 2:142–152.

[3] Ramaswamy S., et al. 2001. Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS*, 98(26):15149–15154.

[4] Su A., et al. Large-scale analysis of the human and mouse transcriptomes. *PNAS*, 99(7):4465–4470.

[5] Tothill RW., et al. 2005. An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. *Cancer Research*, 65(10):4031–4040.

Inference of Protein-Protein Interactions: An Evolutionary Approach

Janusz Dutkowski,[1] Jerzy Tiuryn[1]

1 Introduction

Large amounts of protein-protein interaction (PPI) data generated by high-throughput experimental techniques can provide new insights into the cellular system organization and processes. However, the coverage of interactomes of most model organisms is still low and currently available data is contaminated by false-positive measurements inherent to high-throughput screens. Complementary computational techniques have proven useful in filtering noise and predicting missing interactions [4].

We propose a new method for inferring PPIs which applies an evolutionary model to integrate diverse experimental data from multiple organisms and assign a probability value to each possible interaction. We apply this framework to infer protein interactions in yeast, worm and fly. The interactions identified by the proposed method show strong support in the indirect Gene Ontology (GO) evidence and match many interactions extracted from known protein complexes.

2 Methods

We developed EPPI, a probabilistic framework for predicting protein-protein interactions in multiple species. The method utilizes an extended version of a Bayesian model of protein network evolution presented in [1] which takes into account phylogenetic history of each protein family and the probability of interaction loss or gain during protein duplication or speciation. We extend this model to integrate diverse experimental data from multiple species with different confidence levels (see Fig. 1). Pearl's message passing algorithm [3] is applied to efficiently determine the posterior probability of interaction for each pair of extant proteins.

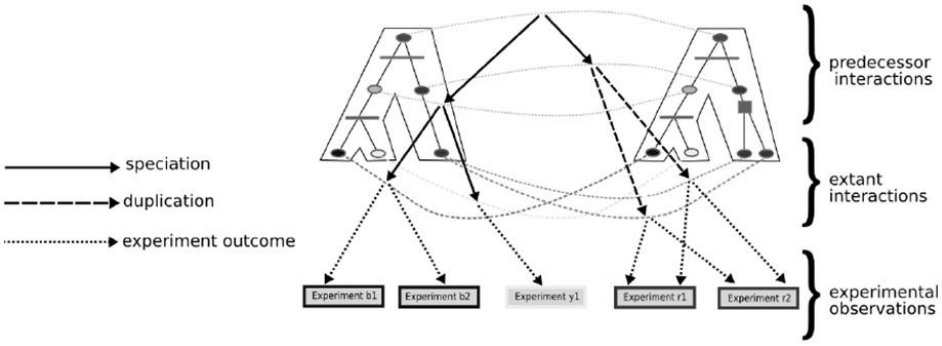


Figure 1: Bayesian tree model of evolution of interactions between members of two protein families for three species: **b**, **y** and **r**. Two reconciled trees for the considered families together with putative protein interactions at each level of evolution are shown. For each species we have a certain number of experimental datasets: two for **b** and **r** and one for **y**. We associate a random variable with each putative protein interaction. Solid arrows indicate dependences between random variables which come from speciation events. Similarly, dashed arrows indicate dependences associated with duplications events. Finally, dotted arrows represent an interface between the true interactions in extant species and the observed experimental evidence.

[1]Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland. Email: {januszd, tiuryn}@mimuw.edu.pl

3 Results

We apply EPPI to infer protein interactions in *S. cerevisiae*, *D. melanogaster* and *C. elegans*. Using a recently proposed scoring scheme [5], we comprehensively assess biological significance of inferred interactions based on biological process (BP), molecular function (MF), and cellular component (CC) assignments in Gene Ontology (see Fig. 2). We also utilize a competing scoring procedure to estimate the rate of true positive and false positive predictions using the MIPS intracomplex interactions as a gold standard. We determine the number of predicted interactions in which both proteins are part of the same yeast complex, as opposed to the number of predictions in which the two proteins are assigned to different sub-cellular localization categories. We conclude that our method yields biologically relevant interactions, both in terms of the GO assignments and in terms of identifying protein pairs present within known complexes. We have found that networks comprised of top EPPI predictions significantly outperform the input datasets used for training. Our method also performs favorably to the domain-based approach presented in [2].

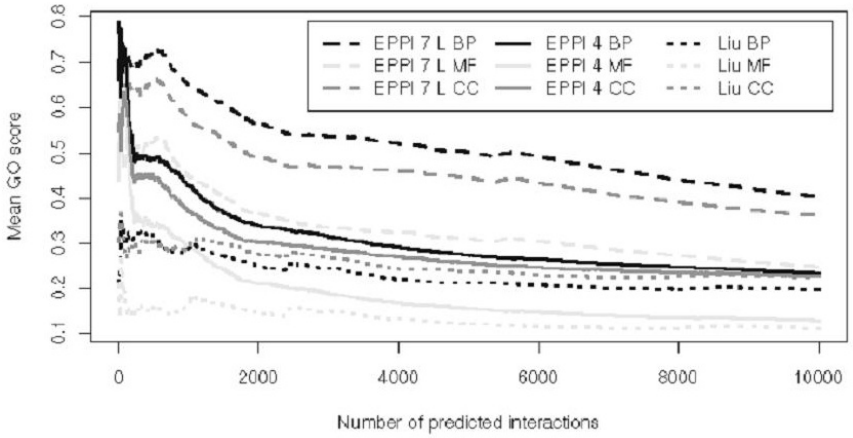


Figure 2: Assessment of predicted yeast interactions using GO functional similarity score (ranging from 0 to 1 with increasing similarity). The similarity of GO annotations for each pair of interacting proteins is measured in each ontology (BP, MF and CC). The interactions are ranked by their probabilities and the average score for the top n predictions is shown. EPPI 4 and EPPI 7 L versions are based on different input datasets and different confidence levels. EPPI 4 uses the same four datasets and uniform confidence levels as in Liu et al [2]. EPPI 7 L uses three additional input datasets and dataset confidence levels derived from literature.

Acknowledgments. This work was supported by the Polish Ministry of Science grants No 3 T11F 021 28 and PBZ-MNiI-2/1/2005.

References

[1] Dutkowski, J., Tiuryn, J. 2007. Identification of functional modules from conserved ancestral protein-protein interactions. *Bioinformatics*, 23:149–158.

[2] Liu, Y., Liu, N., Zhao, H. 2005. Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, 21:3279–3285.

[3] Neapolitan, R.E. 2003. *Learning Bayesian Networks*. Prentice Hall.

[4] Shoemaker, B.A., Panchenko, A.R. 2007. Deciphering Protein Interactions. Part II. Computational Methods to Predict Protein and Domain Interaction Partners. *PLoS Comput Biol*, 3(4):e43.

[5] Schlicker, A., Domingues, F., Rahnenfuhrer, J., Lengauer, T. 2006. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, 7:302.

Integrative Analysis of Transcriptome and Genomic Aberration Map in Cancer

Xing Yi Woo,[1] Edison T. Liu,[2] Guillaume Bourque[3]

1 Introduction

It is well known that copy number alterations in the genome also results in genetic instability and modify gene expression and functions that contributes to tumor progression [1]. Array comparative genomic hybridization (array CGH) has enabled high-resolution and genome-wide detection of copy number variations in the cancer genome [2]. Through comprehensive characterization of the human cancer cell transcriptome, novel transcripts or alternative transcript variants can be identified [3]. In addition, the genome-wide identification of transcription factors binding sites and regulated genes describes the transcriptional activity across the genome in cancer [4]. Hence, an integrative analysis of the transcriptome, gene expression, binding sites and copy number variation maps can possibly lead to the identification of transcripts and genes resulting in cancer progression [5], and thus shedding light into the gene regulation network in cancer.

This poster presents the identification of transcripts, regulated genes and binding sites for different copy number deletions and amplifications for the breast cancer cell line, MCF7, treated with estrogen. The ultimate goal is to develop a systemic and integrative analysis platform for these genome-wide datasets, for the purpose of understanding the transcription regulation pathway in breast cancer cells.

2 Results

The copy number aberrations were obtained from a high-resolution, genome-wide, array CGH profile. The comprehensive transcriptome map of the breast cancer cell line (MCF7), treated with estradiol, was obtained from the GIS-PET technology [3]. A brief analysis shows that 48.8% of the transcripts in the entire transcriptome library have more that 90% overlap with known genes, while 10.5% of the transcripts are identified to be novel transcripts, possibly containing novel genes.

Transcriptome and genomic aberration maps. Figure 1 shows that about 22% of the MCF7 genome is partially or completely deleted, while about 15% of the genome is amplified. Combining the known genes and the transcriptome map with the array CGH data, most of the known genes and transcripts are in the normal (2 copies) and the partially deleted (-1 copy) region. There is also an increase in the number transcripts relative to known genes with copy number.

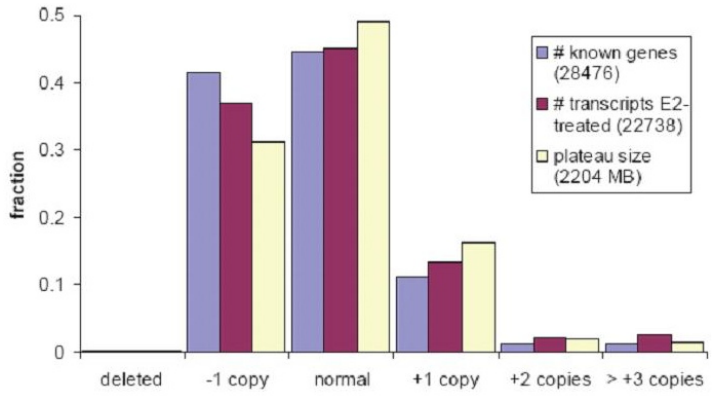


Figure 1: Distribution of copy number variations for MCF7 treated with estradiol, and the distribution of the known genes, transcripts, and the genomic span for the respective deleted and amplified regions.

[1]Computational and Mathematical Biology, Genome Institute of Singapore. Email: wooxy@gis.a-star.edu.sg
[2]Cancer Biology and Pharmacology, Genome Institute of Singapore. Email: liue@gis.a-star.edu.sg
[3]Computational and Mathematical Biology, Genome Institute of Singapore. Email: bourque@gis.a-star.edu.sg

Regulated genes and genomic aberration maps. Using the list of regulated genes obtained from microarray gene expression data [4], an increase in the proportion of regulated genes relative to known genes with copy number is observed, which is consistent with the transcriptome analysis above.

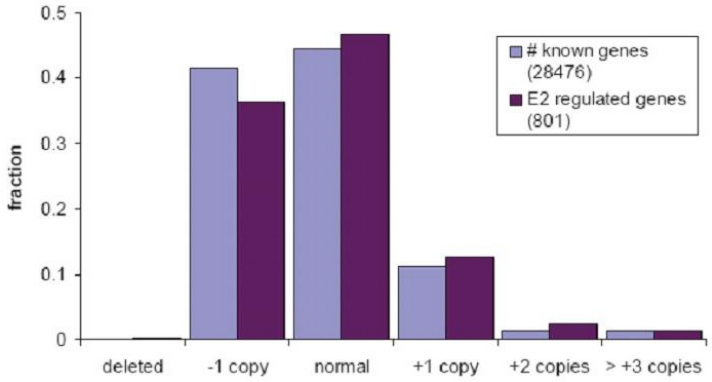


Figure 2: Distribution of known genes and regulated genes for the respective deleted and amplified regions.

Binding sites and genomic aberration maps. The genome-wide identification of estrogen receptor (ER) binding sites describes the transcriptional activity following estrogen treatment [4]. Relative to the genome distribution across the aberrated regions, there is an increase in the proportion of the binding sites with copy number (Figure 3).

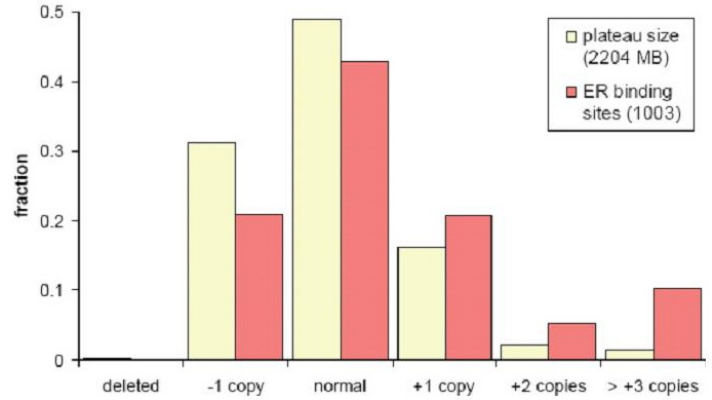


Figure 3: Distribution of genome span and binding sites for the respective deleted and amplified regions.

The combined analysis of the various genome-wide datasets can lead to the understanding of copy number effects in gene regulation in breast cancer with estrogen treatment. This motivates the further development of an integrated and systemic analysis platform for these genome-wide datasets for the understanding of the effects of copy number variations in gene regulation in cancer, and the development of gene regulatory network in breast cancer.

References

[1] Pinkel, D. and Albertson, D. G. 2005. Array comparative genomic hybridization and its applications in cancer. *Nature Genetics*, 37:S11–S17.

[2] Lockwood, W. et al. 2005. Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *European Journal of Human Genetics*, 14:139–148.

[3] Ruan, Y. et al. 2007. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETS). *Genome Research*, 17:828–838.

[4] Lin et al. 2007. Whole-genome cartography of estrogen receptor a binding sites. *PLos Genetics*, 3:e87

[5] Ionita, I. et al. 2006. Mapping tumor-suppressor genes with multipoint statistics from copy-number-variation data. *American Journal of Human Genetics*, 79:13–22.

Discovery of Novel Relationship Among Single Nucleotide Polymorphisms, Alternative Splicing Events and Tumor

Fang Rong Hsu,[1] Wen Chun Lo[2]

1 Introduction

In mankind's genetic difference, the 90% of the diseases are the genetic mutation caused by Single Nucleotide Polymorphism (SNP) [8, 1]. Also, SNP has high stability according to evolution. It has hereditary characteristics. Recent genome-wide analyses of alternative splicing indicate that 40%–60% of human genes have alternative splice forms [3]. The analytic results also believe alternative splice forms are one of the most significant components of the functional complexity of the human genome. At the same time, the cancer of the first of domestic ten major causes of the death. It is also caused by a series of gene mutation [9].

For this reason, we attempt to find the relationship among SNPs, alternative splicing events and tumor.

2 Materials

We used some datasets in our methods. The datasets are as follows. The Avatar Database (A Value Added Transcriptome Database) contains the information of the ESTs of the human sequences and alternative splicing events [4]. (The Avatar website, http://avatar.iecs.fcu.edu.tw/). The EST sequences were downloaded from NCBI (National Center for Biotechnology Information). (The NCBI website, ftp://ftp.ncbi.nih.gov/repository/dbEST/gzipped/dbEST.reports.date.no.gz). The EST sequences in dbEST database has nearly 7.6 million human ESTs. Human genomic sequences were retrieved from NCBI. (The human genomic sequences, build 35, ftp://ftp.ncbi.nih.gov/genbank/genomes/). In addition, the dbSNP database contains the relevant SNP information from NCBI. (The relevant SNP information, build 125, ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/chr_rpts/). The histological information was provided by NCI-CGAP (The Cancer Genome Anatomy Project) Library database. (The NCI-CGAP Library database, ftp://ftp1.nci.nih.gov/pub/CGAP/Hs_LibData.dat).

3 Methods

First, we used Mugup [7, 5] to align ESTs to human genome. It can get the detailed result include mismatch or gap position. We also identified each SNP which may place on exon, using the SNP location data and Avatar exon boundary data. And we get the SNP data which was supported by ESTs. According to the SNP location and alternative splicing events in addition, we distinguish the relationship of SNP and alternative splicing events. Therefore, in finding the relationship between SNP and alternative splicing events, ESTs were divided into four pools: SNP and isoform 1, General nucleotide and isoform 1, SNP and isoform 2, and General nucleotide and isoform 2.

Besides, we found seven million ESTs from 8872 libraries were categorized into 47 tissues and three types of histology, normal, tumor and unknown. In finding the relationship between alternative splicing events and tumor, ESTs were divided into four pools: isoform 1 and tumor, isoform 2 and tumor, isoform 1 and normal, and isoform 2 and normal. In finding the relationship between SNPs and tumor, ESTs were divided into four pools: isoform 1 and tumor, isoform 1 and tumor, isoform 2 and normal, and isoform 2 and normal.

[1]Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan. Email: frhsu@fcu.edu.tw

[2]Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan. Email: m9510587@fcu.edu.tw

Finally, we use Fishers exact test to confirm the result and find the significant information. The Fishers exact test divided left tail of P-value by right tail of P-value as confidence C. The P-value of C was smaller than 0.05, were suggested as certain information.

4 Results

First, we found alternative splicing events which may be caused by SNPs. There were 1,624 such SNPs. In addition, we found 579 cancer-specific alternative splicing events. We also tried to find the relationship between SNPs and tumor. We have found 21,619 cancer-related SNPs.

Finally, we further analysis relationship among SNPs, alternative splicing events and tumor. Among tumor-specific alternative splicing events, some of them may be caused by SNPs. We have identified 36 such SNPs. These SNPs may cause alternative splicing events and further result in tumor

References

[1] Abdel Aouacheria, Vincent Navratil, Ricardo Lopez-Perez, et al. 2007. In silico whole-genome screening for cancer-related single-nucleotide polymorphisms located in human mRNA untranslated regions. *BMC Genomics*, 8:2.

[2] Chia Yang Cheng. 2004. *Discovery of Tumor-Specific Alternative Splicing Sites*. Master's thesis. Asia University.

[3] F. R. Hsu, H. Y. Chang, Y. L. Lin, et al. 2004. Genome-wide alternative splicing events detection through analysis of large scale ESTs. *IEEE 4th Bioinformation Symposium on Bioinformatics and Bioengineering* (BIBE'2004), pp. 310–316.

[4] Fang Rong Hsu, Hwan-You Chang, Yaw-Lin Lin, et al. 2005. AVATAR: A database for genome-wide alternative splicing event detection using large scale ESTs and mRNAs. *Bioinformation*, 1(1):16–18.

[5] F. R. Hsu and J. F. Chen. 2003. Aligning ESTs to genome using multi-layer unique markers. *IEEE Computational Systems Bioinformatics Conference 03*, pp. 564–567.

[6] Hsien Chun Lin. 2004. *Discovery the Relationship between Single Nucleotide Polymorphisms and Alternative Splicing Events*. Master's thesis. Asia University.

[7] L. Florea, G. Hartzell, Z. Zhang, et al. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Research*, 8(9):967–974.

[8] L. Y. Chen, S. H. Lu, E. S. Shih, and M. J. Huang. 2002. Single nucleotide polymorphism mapping using genome-wide unique sequences. *Genome Research*, 12(7):1106–1111.

[9] Qiang Xu and Christopher Lee. 2003. Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Research*, 31(19):5635–5643.

Role of Interaction Databases in Studying Cross-Talks Between Pathways

Arathi Raghunath,[1] Sangjukta Kashyap,[1] Usha Mahadevan,[1] Jignesh Bhate,[1] Pratap Dey[1]

1 Introduction

Bi molecular interaction database constitute a major source of data for understanding signaling pathways or in the course of drug discovery and other biologically relevant studies. NetPro^{TM} is a bi molecular interaction database containing protein-protein, protein-small molecule and small molecule-small molecule interactions. Cross talks between pathways form an important link in understanding interactions in the physiological context. This analysis highlights the utility of interaction databases in understanding cross-talk between pathways.

2 Method and Discussion

Ligands form the starting points of any receptor mediated pathway. In order to understand cross-talks between pathways, we looked at interactions in NetPro^{TM} triggered by different ligands having different effects on a common downstream target. Interactions in NetPro^{TM} brought to light an interesting fact that Insulin and Angiotensin II (AGT II) via their respective receptors can induce activity of PI3K, AKT and ERK1/2. Further evaluation of the interactions highlighted the fact that prior treatment with AGT II for 5 minutes would inhibit insulin-mediated activation of PI3K, AKT and ERK1/2. These data suggest that insulin and angiotensin by themselves activate the same pathway but when present together, angiotensin would interfere with the insulin signaling.

Analysis of the data in NetPro^{TM} along with their kinetic details, for the reasons for this differential regulation, brought out the following facts:

- AGT II increases the level of intracellular calcium as early as 40–60 secs
- Calcium can increase the activity of Protein kinase C as early as 1 min with the maximum activation around 3 mins.
- Protein kinase C can phosphorylate INSR on Ser/Thr residues and inhibit Tyr phosphorylation of INSR and hence its activity.

The above mentioned observations project the following physiological scenario: Insulin through insulin receptor/PI3K/AKT pathway would increase the activity of ERK1/ERK2 while AGT II through its Gprotein coupled receptor would increase levels of calcium and follow the Protein kinase C pathway to increase activity of PI3K/AKT and MAPK. Prior addition of AGT II would inhibit Tyr phosphorylation and activity of INSR and hence any further addition of INS might not have any downstream affect.

AGT has been known to be involved in inducing insulin resistance. This could be the pathway in which it induces resistance to insulin. These data also to some extent justifies the use of ACE inhibitors in inducing insulin sensitivity and in diabetes related complications like diabetic retinopathy, nephropathy etc [1]. Reports of the possible role of protein kinase C in insulin resistance associated disorders might be a further proof for this pathway in insulin resistance [2]. This analysis highlights the role of interaction databases and the contrast information in highlighting the data from various sources enabling one to come to meaningful conclusions. Thus interactions databases form a useful platform, enabling pathway analyzing and cross talks between pathways based on collated data from various sources.

[1]Molecular Connections Pvt. Ltd., Kandala Mansions, 2/2, Kariappa Road, Basavangudi, Bangalore 560004, India. Email: pratap@molecularconnections.com

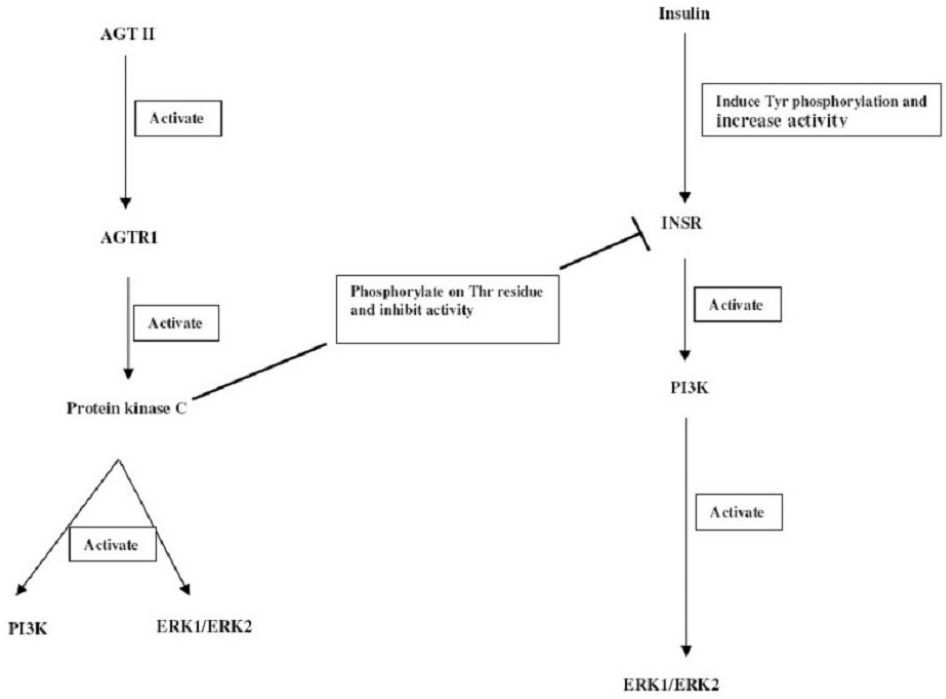


Figure 1: Schematic representations of the networks highlighting the cross talk between angiotensin and insulin signaling.

References

[1] Koike, N., Takamura, T., Kaneko, S. 2007. Induction of reactive oxygen species from isolated rat glomeruli by protein kinase C activation and TNF-alpha stimulation, and effects of a phosphodiesterase inhibitor. *Life Sci*, 80(18):1721–1728.

[2] Smith, DH. 2007. Fixed-dose combination antihypertensives and reduction in target organ damage: Are they all the same? *Am J Cardiovasc Drugs*, 7(6):413–422.

AVATAR II: An Alternative Splicing Database Using Three Alignment Tools[1]

Fang Rong Hsu,[2] Dung-Lin Hsieh,[3] Shao-Peng Yeh[4]

1 Introduction

Alternative splicing (AS) is an important post-transcriptional process and one of the most significant elements leading to increase protein functional complexity. Recent genome-wide analysis of alternative splicing indicate that 40–60% of human genes have alternative splice forms, suggesting that alternative splicing is one of the most significant components of the functional complexity of the organism genome.

Recent alternative splicing database have two major problems: one is the species are too few in AS database that can not to satisfy users, and on genomic mapping processing, most AS database only use single genomic mapping tool, therefore will get lower reliability.

We perposed AVATARII (A value added transcriptome database version 2.0) through an analysis of large scale ESTs. Addition, there are about thirteen species (Figure 1) such as homo sapiens, rattus norvegicus, mus musculus, ..., etc. AVATARII promoted the utilization ratio of ESTs by using multiple genomic mapping tools. In addition, we made the improvement on presenting at AS events.

2 Method and Features

AVATAR [7] compare to ASAPII [2] which used UniGene [5] to cluster EST to assemble a whole gene. However, we already knew UniGene Cluster has not certainly considered Exons/Introns boundary characteristic [1]. It may cause the EST have the wrong result of clustering.

We proposed Mugup(Multi-layer genome-wide unique marker positioning) [3] which is much faster than common tools such as SIM4 [6] and BLAT [4]. In Avatar project, We use Mugup to align ESTs to genome. Mugup use the multi-layer unique markers alignment method by which we can extensively reduce the time required, however, no decrease of specificity and sensitivity. And when aligning ESTs to genome, we also can get the detailed result include mismatch or gap position and the different nucleotide between ESTs and genomic sequence.

AVATARII promoted the utilization ratio of ESTs by using Mugup, Gmap [8] and SIM4 which is described as following. Because each kind of genomic mapping tool be designed with different algorithm, therefore respectively has the advantages and defects in the genomic mapping processing. First, we adopt ESTs that have good mapping score better than a given threshold by using Mugup and Gmap. Finally, we use SIM4 to process ESTs that have lower mapping score in Mugup and Gmap. Take rattus norvegicus as the example, we promoted approximately 2% to be possible the amount of ESTs (Figure 2).

We used simple diagram to express the complex AS event (Figure 3), substitutes for the before edition complicated method of portrayal. In addition, users can request specify AS event to watch some AS event diagram.

References

[1] Burge CB and Karlin S. (1998) Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.*, 8:346–354.

[2] Christopher Lee, Namshin Kim, 2006. ASAPII, `http://bioinformatics.ucla.edu/ASAP2`.

[3] Fang Rong Hsu, Wei-Chung Shia. et al. Single nucleotide polymorphism mapping using multi-layer unique markers. *Journal of Computers*, 18(3):???–???, 2007.

[1] AVATAR II: `http://140.134.26.162/AvatarII/Home.php`

[2] Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan. Email: `frhsu@fcu.edu.tw`

[3] Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan. Email: `M9525023@fcu.edu.tw`

[4] Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan. Email: `M9602012@fcu.edu.tw`

[4] Kent W. (2002) BLAT: The BLAST-like alignment tool. *Genome Res.*, 12:656–664.

[5] NCBI UniGene database, `http://www.ncbi.nlm.nih.gov/UniGene`.

[6] Pidoux AL, Richardson W, Allshire RC. (2003) Sim4: A novel fission yeast kinetochore protein required for centromeric silencing and chromosome segregation. *J Cell Biol.*, 161(2):295–307.

[7] Hsu FR, Chang HY, et al. (2005) AVATAR: A database for genome-wide alternative splicing event detection using large scale ESTs and mRNAs. *Bioinformation*, 1(1):16–18.

[8] Thomas D. Wu and Colin K. Watanabe. (2005) GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinfomatics*, 21(9):1859–1875.

Organism	Exon Skipping	3' site Splice	5' site Splice	Mutual Exclusion	Intron Retention	Total
Homo sapiens(HS)	7728	4292	4085	176	1985	18266
Mus musculus(MM)	3983	4076	5139	251	2437	15886
Rattus norvegicus(RN)	589	614	699	21	656	2579
A.thaliana(AT)	22	191	266	1	191	671
C.elegans(CE)	50	96	121	5	153	425
Macaca Mulatta(MUM)	19	27	45	0	13	104
Anopheles gambiae(AG)	53	96	78	5	56	288
A.mellifera(AM)	25	40	39	3	29	136
Bos taurus(BT)	1284	1261	975	73	1149	4742
Canis familiaris(CF)	277	370	318	17	148	1130
D.melanogaster(DM)	210	315	363	15	533	1436
Danio rerio(DR)	371	753	991	24	534	2673
Gallus gallus(GG)	811	959	1359	74	925	4128

Figure 1: All alternative splicing events about thirteen species on AVATARII.

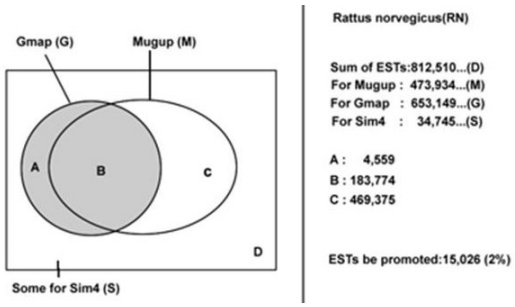


Figure 2: Using multiple genomic mapping tools to promote the utilization ratio of ESTs.

Figure 3: The diagram to express all AS events on AVATARII website.

Large-Scale Comparative Studies in GPFlow

Lawrence Buckingham,[1] Xin-Yi Chua,[1] James M. Hogan,[1]
Paul Roe,[1] Jiro Sumitomo[1]

1 Introduction

Sequence data is typically analyzed through a pipeline of tools, perhaps to align sequences and search for motifs. Tool pipelines are either realized manually or through some kind of script or workflow system. The explosive increase in the number of genomes available has made single sequence analyses almost obsolete. Bioinformaticians now wish to compare and analyze multiple versions of similar sequences, and the greater statistical significance afforded by automated comparisons is vital to scientific investigation.

This work describes recent extensions to the GPFlow scientific workflow system, previously reported in [1]. The system supports interactive experimentation, automatic lifting of computations from single-case to collection-oriented computation, and automatic correlation and synthesis of collections.

GPFlow provides an interactive web-based workflow environment which allows the user to construct workflows from scientifically meaningful components without programming. The system implements a structured data flow model, in which a cumulative data structure is created over the lifetime of a computation. A GPFlow workflow presents as an acyclic data flow graph, yet provides powerful iteration and collection formation capabilities.

2 Data Model

A GPFlow workflow consists of a sequence of *Component* objects organized into an acyclic data flow graph. The restriction to acyclic graphs supports the construction of a simple data structure that captures the entire execution history of the workflow. Each Component acts as a placeholder for a collection of *Processor* objects, the *result set*. A Processor is the fundamental data storage and computational unit in a workflow. When a Component is created it is attached to the type definition of a specific Processor subclass from which the Component derives the name and type of each of its input and output channels.

A Processor plays two roles. As an active object it provides a method called *DoWork* which performs a computational task. As a data capsule it acts as a process activation record, preserving input and output values in publicly accessible data fields, thereby recording the execution outcome of the component applied to a single set of input values. Input and output fields are designated by annotating the fields of a Processor subclass with Input and Output attributes. These attributes are queried by the encapsulating Component to discover its input and output channels.

In addition to a result set, each Component maintains a named list of user input collections, one for each input channel. When a workflow is constructed, connections are established between input channels and the output channels of other Components: an input channel connected in this way is said to be *bound*. Any input channels that are not bound are said to be *unbound*: they will source their values from the user input collection of the same name.

Each Processor output field implicitly defines an output collection, which consists of the array of values obtained by iterating over all Processor records in the associated Component's result set and selecting the value of the corresponding field from each Processor object.

A Component becomes eligible to run when all collections attached to its input fields have been populated. At this time the input collections are queried to create a set of partially populated Processor records, in which all input fields have been assigned values but output fields retain their default values. The resulting incomplete activation records are queued for execution. Subsequently, each Processor is invoked to populate its output fields, thus completing the result set for the Component.

3 Data-Driven Iteration

At the most basic level, a Component iterates over the Cartesian product of its input collections and queues one Processor for each combination. This works well if the data flow graph is a simple pipeline

[1]Microsoft QUT eResearch Centre, Queensland University of Technology. GPO Box 2434, Brisbane QLD 4001. AUSTRALIA. Email: {l.buckingham, x.chua, j.hogan, p.roe, j2.sumitomo}@qut.edu.au

or strictly divergent tree structure, but if the data flow graph contains a fork-merge subgraph the simple iteration model breaks down by introducing spurious computations that could never have occurred if the user inputs were locked in manually. We remedy this situation by introducing *key-based correlation*.

Key-based correlation exploits the fact that every data value in a user input collection has a unique and well-defined address, from which the originating component, collection and position within that collection can be deduced. We use the user input addresses to form a key for each result in the system. The key of a user input value is its own address. Any value derived, directly or indirectly, from a user input value, contains the address of that value as part of its key. Thus a key is a list of user input addresses which encodes the provenance of each derived value.

We consider two user input addresses to be *comparable* if and only if they originate in the same channel of the same component. Two keys k_1 and k_2 are said to be correlated if and only if

- No address a_1 in k_1 is comparable to an address a_2 in k_2, or
- For every address a_1 in k_1 that is comparable to an address a_2 in k_2, $a_1 = a_2$.

That is, two values are correlated if and only if they derive from completely distinct lineages, or in the case that they are derived from overlapping sets of channels, they are derived from the same value in each of those channels. To preserve structural integrity of the iterated workflow, a Processor is only queued for execution when its input values are mutually correlated. This also removes the need for user intervention to specify the correlation mode.

4 Collection Formation: Aggregation and Key-Slicing

GPFlow provides two ways to form a collection: *aggregation* and *key-slicing*. An example of aggregation is where we wish to merge the elements of two parallel arrays to form a single array of 2-dimensional vectors. An example of key-slicing is where we wish to gather all values produced by a component to perform some synthesis or summarizing operation.

If an input field has type "Array of T" for some type T, it may be connected to one or more output channels, provided their types are T or "Array of T". When a value is assembled for such an input field, a single sequence is constructed. This sequence contains all constituent elements of the nominated antecedent output channels, subject to the key correlation criterion described above. This extension alone is sufficient to enable aggregation.

Key-slicing is based on the observation that the collected keys of the result set generated by a component form a discrete hypercube, with dimensionality defined by the set of keys that index the elements of the result set. An output value is associated with each point in this hypercube. If we remove a key field, we project onto a hypercube of lower dimension, each point of which indexes a collection of values, namely those distinguished by the value of the key removed. Intuitively, we take a slice through the result cube. To implement this in GPFlow, we permit the user to nominate one or more unbound input variables to be removed from the key for a particular input channel. Sliced inputs may belong to the component that contains the output channel or to any of its antecedents. Any sliced input values are selectively ignored when the correlation test is applied during input value assembly.

5 Conclusions

The problem of managing collections arising in comparative studies is fundamental to post-genome bioinformatics. This work introduces a novel key based approach to tracking data tuples, ensuring correctness and enabling convenient selection from the available Cartesian product of data vectors. In addition to the guarantee of result integrity, the approach provides a ready made platform for provenance tracking and reporting. Above all, the system supports automated lifting of computations, allowing the user conveniently to prototype singleton computations and routinely apply them to full scale data sets.

References

[1] A. Rygg, P. Roe, O. Wong. GPFlow: An Intuitive Environment for Web Based Scientific Workflow. In: *5th International Conference on Grid and Cooperative Computing Workshops*, pp. 204–211, IEEE Computer Society, 2006.

Is Transcription Factors Mediated Gene Regulation Hard Wired? A Microarray-Based Statistical Estimate

Vincent Piras,[2] Alessandro Giuliani,[1] Naoki Fujikawa,[2] Masaru Tomita,[2] Kumar Selvarajoo,[2,*] Masa Tsuchiya[2,*]

1 Summary

Understanding of dynamic control of gene regulatory networks (GRNs) is a prime challenge in molecular biology. A specific GRN is underpinned by dynamical interaction of transcriptionalregulatory systems whose starting point is the binding of a transcription factor (TF) to the promoter region of the gene to be expressed. Thus the binding of a TF to its specific target is considered as the elementary component of GRNs [1, 2, 6]. Even if it is widely accepted that gene regulation involves players different from TFs, such as micro-RNAs and post-transcriptional regulation by means of m- RNA degradation and translational repression [3], the TF based model is still considered as the main responsible of the selective activation of different GRNs. The DNA sequences specifically recognized by each TF are now known for hundreds so giving the possibility to check the feasibility of the construction of a cell regulation map in terms of TFs sharing. In other words, if TF based regulation is the main responsible for the generation of preferential gene regulation circuits, we should expect genes involved in same GRNs to share higher proportion of common TFs than expected by chance.

We challenged this hypothesis by performing a statistical experiment on a microarray dataset referring to the innate immune response to LPS stimulus regulating expression of pro-inflammatory cytokines such as *Tnf*, *Il6*, *Il12*, and interferons such as *Ifn-α* and *Ifn-β* via Toll-Like Receptor 4 (TLR4) pathways. We compared set of genes that were observed to be strongly co-regulated against randomly extracted groups of genes and calculated, at a population level, specificity and commonality of TFs binding sites determined from sequence analysis of promoter regions and database search.

We show only some (15%) groups of co-regulated genes, such as genes co-regulated with *Tnf*, *Nfkbia*, *Irg1*, *Cxcl2*, etc. having TFs *commonality* higher than random populations. However, majority of groups of co-regulated and random genes resulted to be practically identical in terms of TFs *commonality*. This shows, as for micro RNAs [5], that static analysis of TFs sharing cannot systematically determine the fact that genes tend to be involved in the same regulation pathways and that organization of the entire genome expression regulation is far from a rigid 'hard wired' system, and indicates the importance of dynamic activity of TFs and other regulatory systems such as non coding RNAs mediated regulations.

2 Materials and Methods

Promoter regions and Transcription Factors binding sites databases. We used DBTSS database (http://dbtss.hgc.jp) that contains the [-1000; +200] promoter regions for *mus musculus*. The TF binding sites database we used is the commercial version of Transfac database which contains information for binding sites of *mus musculus* (http://www.biobase-international.com).

Microarray data. The microarray data we based our analysis upon, contains expression levels for 22690 ORFs, and refers to the response of *mus musculus* macrophages to Lipopolysaccharide (LPS) stimulus triggering the activation of innate immunity (TLR4 pathway) [4]. The cell response was observed in 3 time points (0, 1 and 4h) and wild type and 3 mutated phenotypes (MyD88 KO, TRIF KO, MyD88/TRIF KO).

Sequence analysis and Database search. Patch algorithm and Transfac database were used over promoter regions of DBTSS database to build a dataset containing TF binding sites information for the different ORFs. We used Fatigo+ web tool (http://babelomics.bioinfo.cipf.es) to extract the TF binding

[1]Istituto Superiore di Sanitá, Environment and Health Department, Rome, Italy.
[2]Institute for Advanced Biosciences, Keio University, Tsuruoka, Japan.
* Email: kumar@ttck.keio.ac.jp, tsuchiya@ttck.keio.ac.jp

sites of the different ORFs, predicted using Match program and Transfac database in the 1kb 5' region of the gene with high quality matrices.

Statistical analysis. Using Principle Component Analysis, we selected 136 genes highly regulated in LSP response. We extracted 136 groups of N genes co-regulated with each of these genes (with high Pearson correlation in gene expression) and calculated binding sites *commonality* (average binding sites commonality of all pairs of genes defined by: Number of TF binding sites shared among a pair of genes/Number of binding sites of both genes) among groups of co-regulated genes, then compared these values with those obtained with 30 groups of N randomly selected genes and calculated 95% confidence interval for statistical significance.

3 Figures and Tables

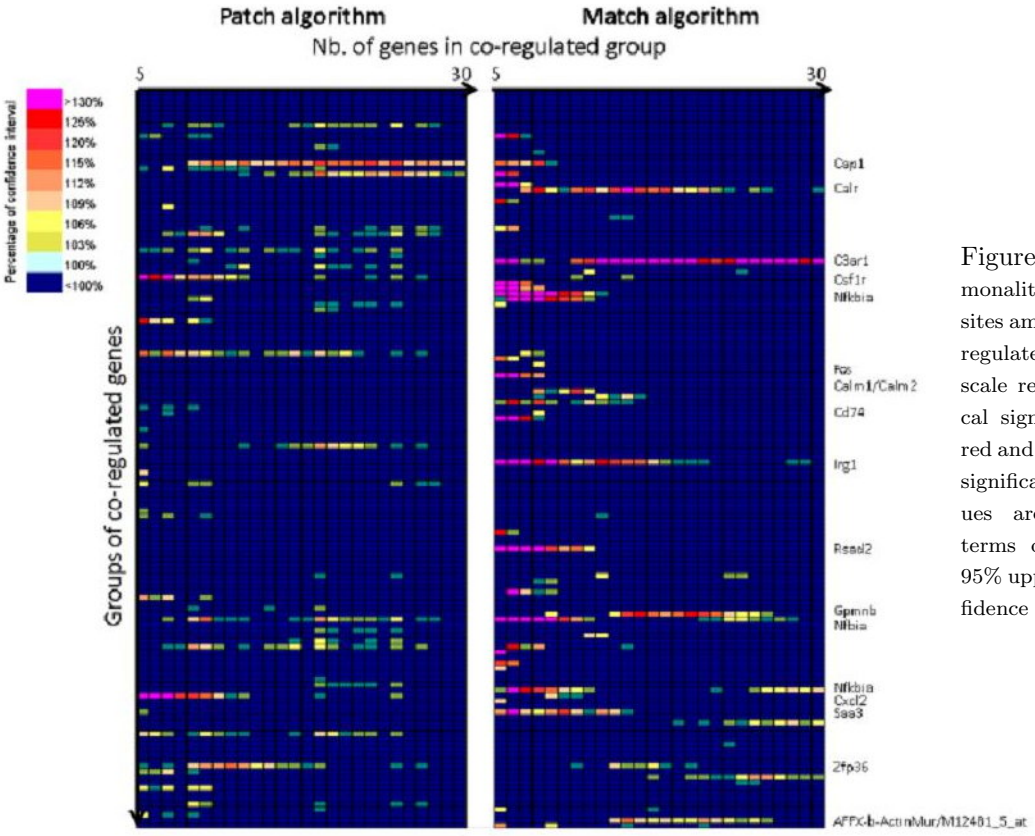


Figure 1: Overall commonality of TF binding sites among groups of co-regulated genes. Color scale represents statistical significance (purple, red and orange for highly significant values). Values are calculated in terms of percentage of 95% upper bound of confidence interval.

References

[1] Bintu, L., Buchler, N.E., Garcia, H.G., et al. 2005. Transcriptional regulation by the numbers: Models. *Current Opinion in Genetics & Development*, 15:116–124.

[2] Bintu, L., Buchler, N.E., Garcia, H.G., et al. 2005. Transcriptional regulation by the numbers: Applications. *Current Opinion in Genetics & Development*, 15:125–135.

[3] Chen, K., Rajewsky, N. 2007. The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics*, 8:93–103.

[4] Hirotani, T., Yamamoto, M., Kumagai, Y., et al. 2005. Regulation of lipopolysaccharide-inducible genes by MyD88 and Toll/IL-1 domain containing adaptor inducing IFN-beta. *Biochemistry and Biophysics Research Communications*, 328:383–392.

[5] Piras, V., Selvarajoo, K., Fujikawa, N., et al. 2007. Statistical Analysis of Gene Expression in Innate Immune Responses: Dynamic Interactions between MicroRNA and Signaling Molecules. *Genomics and Informatics*, 5:107–112.

[6] Wolf, D.M., Eeckman, F.H. 1998. On the relationship between genomic regulatory element organization and gene regulatory dynamics. *Journal of Theoretical Biology*, 195:167–186.

Contrast Interaction Database: A Novel Approach to Study Contextual Relevance of Interactions

Arathi Raghunath, Pratap Dey, Usha Mahadevan, Jignesh Bhate, Sangjukta Kashyap[1]

1 Introduction

Molecular interactions constitute the basis for all physiological processes in a cell. Till recent times, the major afflict faced by the scientific community was the inability to use information about molecular interactions dispersed in scientific literature. The birth of interaction databases catered data in a structured format, which became an essential and easily searchable resource for biologists. Having an easily searchable repository in hand, the scientific communities are now looking for more granular information to validate the data more precisely. In an effort to understand the needs of the scientific community we closely studied the details of molecular interactions. A very common revelation was that all interactions are tightly regulated by certain conditions or factors. The nature of interaction between two interacting molecules varies based on various factors, so much so that we found interesting instances of contrasting interactions between pair of molecules every now and then. Making a conscious effort, we built a pioneering database, "NetPro Contrast Interactions database" based on this concept to bring such information at the disposal of the scientific community which otherwise would not have been easy to retrieve. This study is made to bring contrast interactions into focus and state the importance of capturing such granular information by correlating with biologically relevant instances.

2 Source of data

Contrast Interaction database: We utilized data from "NetPro Contrast Interactions Database" that contains sets of interactions, hand curated from literature, where the Effector has been shown to have contrast effects on the affected molecule or interactors has been shown to have contrast associating tendency. Interactions are supplemented with experimental conditions data and domain, residue details to explain the difference of contrasting effects. We studied a set of interactions in from the database as shown below.

EFFECTOR MOLECULE					
MC Gene Id	EZHsP00092	Symbol	Rho-associated, coiled-coil containing protein kinase	Species	Homo sapiens
AFFECTED MOLECULE					
Cas Id	10102-43-9	Symbol	Nitric oxide		
General Information					
Verb	Inhibits level (indirect)		Increases level (indirect)		
Experimental location and method	species:: Human cell:: AV3 cells condition:: eNOS mediated Nitric oxide production	VS	organ / tissue:: Spinal cord condition:: nNOS mediated Nitric oxide production upon stimulation by Prostaglandin E2		
Reference			Reference		
Interaction id	156616		217038		
PubMed Id	12446767		16188227		

NO has a role in neurotransmitter release and is also implicated to have a role in memory and learning by regulating long-term potentiation (LTP) in the hippocampus. Recent studies also suggest its role in sleepwake cycle. One of the many mechanisms for NO production in brain involves signaling via the principal pro-inflammatory prostanoid, PGE2, as evidenced by the interaction from the database as

[1]Molecular Connections Pvt. Ltd., Kandala Mansions, 2/2, Kariappa Road, Basavangudi, Bangalore 560004, India. Email: sangjukta@molecularconnections.com

shown above (interaction id 217038). PGE2 stimulates NO production in brain stem via the following cascade: PGE2 → EP3 → Rho-kinase → nNOS → ↑NO.

Apart from its role in brain development, NO is heavily involved in smooth muscle relaxation. Nitric oxide as a vasodilator is speculated to be effective in keeping blood vessel plaque free as it reduces the tendency of white blood cells and platelets to aggregate on the walls of the vessel. Endothelial nitric oxide synthase (eNOS) is an important regulator of cardiovascular homeostasis leading to production of nitric oxide (NO) from vascular endothelial cells [PKB → eNOS → ↑NO]. Interestingly, in contrast to nNOS, eNOS is inhibited by Rho-kinase via PKB inhibition, essentially leading to decrease in NO production as shown above (interaction id 156616) [Rho-kinase ⊣ PKB ↛ eNOS ↛ ↓NO].

3 Discussion

Studies suggest that inhibition of Rho-kinase pathway may play a key role in the cardio protective effect on cardiovascular remodeling associated with eNOS. Mita S et al. (2005) showed that treating Dahl saltsensitive hypertensive rats with a specific Rho-kinase inhibitor, Y-27632, significantly ameliorated increased left ventricular weight in the hypertrophy stage. It also effectively inhibited vascular lesion formation, such as medial thickness and perivascular fibrosis. Hence they suggested that Rho-kinase pathway inhibition could be a potential therapeutic strategy for hypertension with cardiac hypertrophy. We realized that the interaction database could be important in suggesting such crucial therapeutic strategy in research. Data from interaction id 217038 may give an insight of the expected effect of nNOS mediated NO production upon Rho-kinase inhibition. Such data would make scientists conscious about the negative phenotypic characteristics that may arise upon target inhibition.

This paper also elaborates on instances of contrasting interactions in the database which could be of use to the research groups in making crucial decisions like target identification, and also in many cases would help normalization of experimental set ups.

References

[1] Matsumura, S., Abe, T., Mabuchi, T., et al. Rho-kinase mediates spinal nitric oxide formation by prostaglandin E2 via EP3 subtype. *Biochem Biophys Res Commun*, 2005, 338(1):550–557.

[2] Ming, X.F., Viswambharan, H., Barandier, C., et al. Rho GTPase/Rho kinase negatively regulates endothelial nitric oxide synthase phosphorylation through the inhibition of protein kinase B/Akt in human endothelial cells. *Mol Cell Biol*, 2002, 22(24):8467–8477.

[3] Mita, S., Kobayashi, N., Honda, T., et al. Cardioprotective mechanisms of rho-kinase inhibition associated with eNOS and oxidative stress-LOX-1 pathway in Dahl salt-sensitive hypertensive rats. *Journal of Hypertension*, 2005, 23(1):87–96.

[4] `http://sulcus.berkeley.edu/mcb/165_001/papers/manuscripts/_638.html`

[5] `http://www.acnp.org/g4/GN401000060/CH060.html`

Finding MicroRNA-mRNA Modules Based on Rule Induction

Dang Hung Tran,[1] Kenji Satou,[2] Tu Bao Ho[3]

1 Introduction

MicroRNAs (miRNA) are a class of small noncoding RNA molecules (20–24 nt), which are believed to participate in down-regulation of gene expressions. They can inhibit their target genes (mRNA) at posttranscriptional process by complementary base pairing. Researchers have devoted many attempts to understand the function of miRNAs in cellular processes more clearly using both experimental and computational methods. Most efforts concentrate on finding miRNAs and their targets [1]. The relationship between miRNAs and their target genes, however, is generally complicated. One target gene could be regulated by several miRNAs and conversely, one miRNA may have several target genes. In order to understand the regulatory mechanism of miRNAs in complex cellular systems, we need to identify the functional modules involved in complex interactions between miRNAs and their target genes. Yoon and De Micheli introduced the concept of miRNA regulatory modules (MRM) in 2005 [8], it was defined as groups of miRNAs and their target genes that are believed to have similar functions or involved in similar biological processes. Nevertheless, the main drawback of their method is that it deals only with miRNA-mRNA duplexes in the sequence level. Another approach, proposed by Joung et al [2], tries to combine multiple information sources to extract the MRMs. This method relies on population-based probabilistic learning, whose result quality depends on many sensitive parameters thus making it unreliable. Recently, we introduced a method based on closed itemset mining for finding coherent MRMs [7]. The method considered only the similarity between mRNA expression profiles and the putative mRNA-miRNA relationships. In this paper, we developed a new method that combines expression profiles of miRNAs and mRNA with miRNA-target gene binding information to discover the MRMs. Our method is based on rule induction, a machine learning technique that can efficiently deal with subgroups discovery. The MRMs, which are found by our methods, consist of highly-related miRNAs and their target genes on aspect of biological functions.

2 Methods

2.1 Approach Overview

The problem can be formulated as follows: given a set of miRNAs (mi_1, mi_2, ..., mi_M) and a set of their target genes (m_1, m_2, ..., m_N). We need to find a set of MRMs, each MRMs can be defined as groups of a subset of miRNAs (mi_{i1}, mi_{i2}, ..., mi_{ik}) and a subset of target genes (m_{j1}, m_{j2}, ..., m_{jl}), where $|ik| \leq |M|$ and $|jl| \leq |N|$.

Figure 1 shows procedural steps of our approach. Firstly, we use a clustering method to divide the set of target genes into clusters. Genes in the same cluster have more similar expression profiles than genes in different clusters. Secondly, for each cluster, we build a decision table by adding a class-column into miRNA binding information table. We then apply the CN2-SD rule induction system [4] to produce a set of miRNA-mRNA regulatory rules. After that we use a filtering procedure to discard uninteresting rules. Only significant rules, which contain the miRNAs with highly correlated expression profiles, are considered. Finally, these rules will be evaluated by using Gene Ontology.

2.2 Datasets

In our experiments, we extracted the expression profiles of miRNAs and mRNAs from the experimental data previously published by Lu et al [5]. This dataset consists of 217 miRNAs and about 16063 mRNAs

[1]Japan Advanced Institute of Science and Technology. Email: hungtd@jaist.ac.jp
[2]Kanazawa University. Email: ken@t.kanazawa-u.ac.jp
[3]Japan Advanced Institute of Science and Technology. Email: bao@jaist.ac.jp

on 89 multiple human cancer samples. The miRNA-mRNA binding dataset was obtained from Krek et al [3]. From two kinds of data, we analyzed the relationships among 121 human miRNAs and 801 mRNAs, which are linked together. Of these 801 mRNA x 121 miRNA possible binding pairs, 4629 pairs with significant binding scores (PicTar's score = 1.0) were used in our experiments.

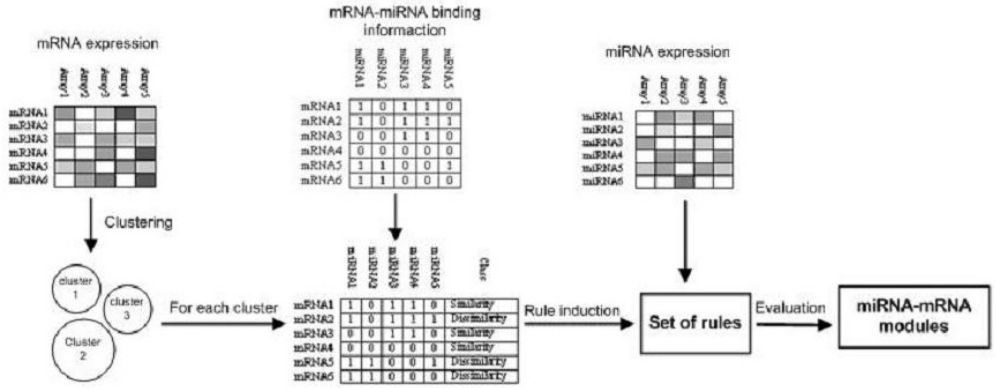


Figure 1: A schematic description of our approach for finding MRMs from predicted target genes and two respective expression profiles of miRNAs and mRNAs.

3 Results and Discussion

We applied our method on the human miRNA datasets as described above. Of 276 potential MRMs found, we keep only a set of 189 coherent MRMs for analyzing after removing some trivial and uninteresting MRMs. Each of above set contains 4.18 miRNAs and 6.74 target genes on average. To validate these modules, we calculated correlation coefficients between miRNAs. So our modules contain not only the similarity between mRNAs expression profiles but also the similarity between related miRNAs. In order to find significant modules with aspect of biological functions, we used the GO:termFinder tool (with a corrected P-value = 0.05) to confirm genes in the same module. This analysis reveals that genes included in the modules have similar biological functions. Moreover, the supporting evidences from literature showed that genes and miRNAs in our modules related to human cancer diseases (e.g. breast cancer) [6]. In future, we plan to apply our method on plant and other animal miRNA genes.

References

[1] Brown J. and Sanseau P. 2005. A computational view of microRNAs and their targets. *Drug Discovery Today: Biosilico*, 10(8):595–601.

[2] Joung G. J., Hwang B. K., Ban W. J., et al. 2007. Discovery of microRNA-mRNA modules via population-based probabilistic learning. *Bioinformatics*, 23(9):1141–1147.

[3] Krek A., Grun D., Poy M. N., et al. 2005. Combinatorial microRNA target predictions. *Nature Genetics*, 37:495–500.

[4] Lavrac N., Kavsek B., Flach P. and Todorovski L. 2004. Subgroup discovery with CN2-SD. *J. Machine Learning Res.*, 5:153–188.

[5] Lu J., Getz G., Miska A. E., et al. 2005. MicroRNA expression profiles classify human cancers. *Nature*, 435:834–838.

[6] Negrini M., Ferracin M., Sabbioni S. and Croce M. C. 2007. MicroRNAs in human cancer: From research to therapy. *Journal of Cell Science*, 120:1833–1840.

[7] Tran D. H., Satou K., and Ho T. B. 2008. Mining microRNA regulatory modules from microRNA-target gene information and expression profile data. In: *6th Asia-Pacific Bioinformatics Conference*, Kyoto, January 14–17, P059.

[8] Yoon S. and Micheli G. D. 2005. Prediction of regulatory modules comprising microRNAs and target genes. *Bioinformatics*, 21(2):ii93–ii99.

Computational Studies on Role of Large Hydrophobic Residues in Proteins

V. Jayaraj,[1] R. Suhanya,[2] M. Vijayasarathy,[3] E. Rajasekaran[3]

1 Introduction

There is lot of work gone into proteins to understand the ultimate truth of real hideous information [1, 2, 3, 4, 5, 6]. To understand the nature of proteins further, the role of large hydrophobic residues in globular proteins are studied here.

2 Methodology

The probable number of large hydrophobic residues in the given window length or number of amino acids is computed using the programs written in C. That is the total number large hydrophobic residues are counted for a given window length. The windows are grouped based on the number of large hydrophobic residues. This grouped number of windows in the given number of large hydrophobic residue is further divided by total number of window in the given species to get the point at which maximum frequency as shown in the figure. Though the window length taken from 5 to 90, only window length 45 is plotted here for discussion. The protein sequences of human, chimpanzee, bovine, dog, rat, mouse, chicken, zebrafish, fruitfly, honeybee, mosquito, roundworm, *A. thaliana*, yeast, *K. lactis*, *E.coli*, Influenza virus, *P. falciparum* and *V. cholerae* are taken from ftp site of NCBI website and analyzed.

3 Results and Discussion

The frequency plot as a function of large hydrophobic residues is shown in Figure 1. Note that the maximum frequency is observed at 26.67% (ie., 12/45; number of large of hydrophic residues/window length) for human. Similarly for yeast it is 28.89% (ie., 13/45). Though more species at different window length are studied, only two of them (Yeast and Human) are shown for window length 45. It is observed that given any length, the number of large hydrophobic residues per unit length is preferably 27%. In other term the globular proteins prefers to have 27% of large hydrophobic residues.

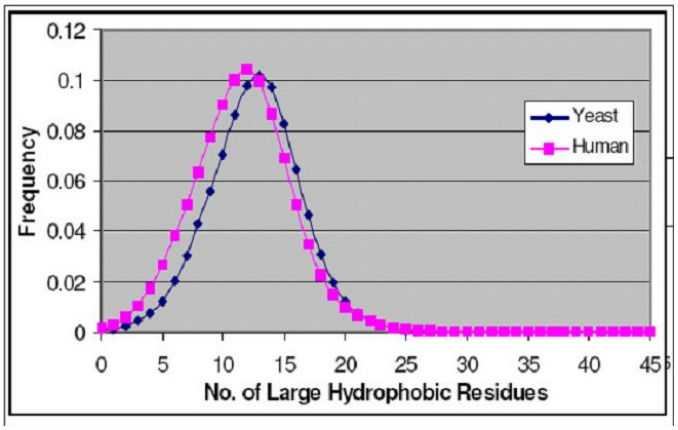


Figure 1: Distribution of windows as a function of large hydrophobic residues for Yeast (blue diamond) and Human (pink square) for window length of 45 amino acids.

[1]Dept of Computer applications, Periyar Maniammai University, Thanjavur 613403, Tamil Nadu, India.
[2]Research Scholar, Bharathidasan University, Trichy 620024, Tamil Nadu, India.
[3]Dept of Biotechnology, Periyar Maniammai University, Thanjavur 613403, Tamil Nadu, India.
* Correspondence: ersekaran@gmail.com

This number of large hydrophobic residues per unit length is observed to be less in heterosexuals. In another study it is observed that the length of proteins higher in heterosexuals. To balance the carbon amount, more number of small hydrophobic residues is added in the proteins. So the length of the animal proteins increases. Globular proteins are expected to follow this distribution profile. It is noticed that more number of large hydrophobic residues in the active site. This increase in the active site is adjusted (decreased) in other portions of the protein to maintain 27% of large hydrophobic residues.

4 Conclusion

One of the interesting observations is that except the active site the other portions of the proteins are maintaining a definite fraction (27%) of large hydrophobic residues that gives local stability all along the sequence or structure. Heterosexual animal show up less number of large hydrophobic residues than fungi or plants. The diseased sequences lack these large hydrophobic residues in total or in some portion along the sequences that leads to malfunctioning of the protein.

References

[1] Chothia C and Lesk AM. 1986. The relationship between the divergence of sequence and structure in protein. *EMBO J.*, 5, 823–826.

[2] Narang P, Bhushan K, Bose S and Jayaram B. 2006. Protein structure evaluation using an all atom energy based empirical scoring function. *J. Biomol. Str. Dyn.*, 23, 385–406.

[3] Srinivasan N, Sowdhamini R, Ramakrishnan C and Balram P. 1991. Analysis of short loops connecting secondary structural elements in proteins. In: *Molecular Conformation and Biological Interactions*, Indian Academy of Sciences, pp 59–73.

[4] Ponnusamy PK and Gromiha MM. 1996. Hydrophobic distribution and spatial arrangements of amino acid residues in membrane proteins. *Int. J. Pept. Protein Res.*, 48, 452–460.

[5] Venkatachalam CM. 1968. Stereo chemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide unit. *Biopolymers*, 6, 1425–1436.

[6] Ramachandran GN and Sasisekharan V. 1968. Conformation of polypeptides and proteins. *Adv. Protein Chem.*, 23, 283–438.

HP Distance Via Double Cut and Join Distance

Anne Bergeron,[1] Julia Mixtacki,[2] Jens Stoye[3]

1 Introduction

The genomic distance problem in the Hannenhalli-Pevzner theory is the following: Given two genomes whose chromosomes are linear, what is the minimum number of inversions and translocations that transform one genome into the other? The first answer to this question was given by Hannenhalli and Pevzner [2] in 1995. Their distance formula for calculating the so-called *HP distance*, denoted by d_{HP}, requires preprocessing steps such as*capping* and *concatenation* and involves seven parameters. In the last decade, different authors pointed to problems in the original formula and in the algorithm given by Hannenhalli and Pevzner. The algorithm was first corrected by Tesler [5]. In 2003, Ozery-Flato and Shamir [4] found a counter-example and modified one of the parameters of the distance formula. Very recently, another correction was presented by Jean and Nikolski [3].

In contrast to this rather complicated distance measure, Yancopoulos et al [6] presented a general genome model that includes linear and circular chromosomes and introduced a new operation called *double cut and join*, or shortly DCJ. Such an operation can be viewed as making up to two cuts in a genome and joining the resulting segments in any order. In addition to inversions and translocations, the DCJ operation also models transpositions and block-interchanges by creating an intermediate circular chromosome that is re-integrated by another DCJ operation. Beside the simple distance computation, the sorting algorithm is also basic and efficient [1].

On this poster, we will show how the rearrangement model considered in the HP-theory can be integrated in the more general DCJ model. Moreover, given the DCJ distance d_{DCJ}, the HP distance can be expressed as

$$d_{HP} = d_{DCJ} + t$$

where t represents the extra cost of not resorting to unoriented DCJ operations. This extra cost can easily be computed by a tree data structure associated to a genome.

2 Components and Oriented Sorting

Let A and B be two linear multi-chromosomal genomes on the same set of N genes. A linear chromosome will be represented by an ordered sequence of signed genes, flanked by two unsigned telomere markers:

$$(\circ, g_1, \ldots, g_n, \circ).$$

An interval $(l, \ldots, r)$ in a genome is a set of consecutive genes or telomere markers within a chromosome; the set $\{l, -r\}$ is the set of *extremities* of the interval; note that $\circ = -\circ$. An *adjacency* is an interval of length 2, an adjacency that contains a telomere marker is called a *telomere*.

Definition 4 *Given two genomes A and B, an interval $(l, \ldots, r)$ of genome A is a* component *relative to genome B if there exists an interval in genome B with (a) the same extremities, (b) the same set of genes, and (c) that is not the union of two such intervals.*

Definition 5 *A component is* oriented *if there exists a DCJ operation acting on its adjacencies and telomeres that reduces the DCJ distance and that does not create circular chromosomes; otherwise it is unoriented.*

Theorem 2 *Given two linear genomes A and B, $d_{HP}(A, B) = d_{DCJ}(A, B)$ if and only if there is no unoriented component.*

[1]Dépt d'informatique, Université du Québec à Montréal, Canada. Email: bergeron.anne@uqam.ca
[2]International NRW Graduate School in Bioinformatics and Genome Research, Universität Bielefeld, Germany. Email: julia.mixtacki@uni-bielefeld.de
[3]Fakultät, Universität Bielefeld, Germany. Email: stoye@techfak.uni-bielefeld.de

3 The HP Distance Formula

Two components are either disjoint, nested, or they overlap on one gene. When they overlap on one gene, we say that they are *linked*. Successive linked components form a *chain*. A chain that cannot be extended to the left or right is called *maximal*.

Given a genome A, we define a tree T as follows: For each chromosome X of A, the linking and nesting relation of the components of X define a forest F_X where components are represented by round nodes and maximal chains of components are represented by square nodes. The root nodes of all trees of the F_X are children of one round node, the root of T. The round nodes of T are painted according the following classification: (1) The root and all nodes corresponding to oriented components are painted *black*. (2) Nodes corresponding to unoriented components that do not contain telomeres are painted *white*. (3) Nodes corresponding to unoriented components that contain one or two telomeres are painted *grey*. An example is given in Fig. 1.

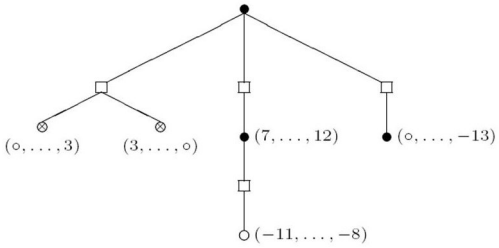


Figure 1: The tree T associated to the genomes $A = \{(\circ, 2, 1, 3, 5, 4, \circ), (\circ, 6, 7, -11, -9, -10, -8, 12, 16, \circ), (\circ, 15, 14, -13, 17, \circ)\}$ and $B = \{(\circ, 1, 2, 3, 4, 5, \circ), (\circ, 6, 7, 8, 9, 10, 11, 12, \circ), (\circ, 13, 14, 15, \circ), (\circ, 16, 17, \circ)\}$ has two grey leaves, one white leaf and one black leaf.

Let T be the tree associated to the components of genome A relative to genome B, and let T' be the smallest subtree that contains all the unoriented components, that is, the white and grey nodes.

Definition 6 *A* cover *of T' is a collection of paths joining all the unoriented components, such that each terminal node of a path belongs to a unique path.*

A path that contains two or more white or grey components, or one white and one grey component, is called a *long* path. A path that contains only one white or one grey component, is a *short* path. The *cost* of a cover is defined to be the sum of the costs of its paths, where the cost of path is the increase in DCJ distance caused by destroying the unoriented components along the path: (1) The cost of a short path is 1. (2) The cost of a long path with just two grey components is 1. (3) The cost of all other long paths is 2. An *optimal* cover is a cover of minimal cost.

Theorem 3 *Given two linear genomes A and B. Then $d_{HP}(A, B) = d_{DCJ}(A, B) + t$ where t is the cost of an optimal cover of T'.*

4 Conclusion

We have given a simpler formula for the Hannenhalli-Pevzner genomic distance equation. It requires only a few parameters that can easily be computed directly from the genomes and from simple graph structures derived from the genomes.

References

[1] A. Bergeron, J. Mixtacki, and J. Stoye. A unifying view of genome rearrangements. In: *Proc. WABI 2006*, LNBI 4175:163–173, 2006.

[2] S. Hannenhalli and P. A. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). In: *Proc. FOCS 1995*, pages 581–592.

[3] G. Jean and M. Nikolski. Genome rearrangements: A correct algorithm for optimal capping. *Inf. Process. Lett.*, 104:14–20, 2007.

[4] M. Ozery-Flato and R. Shamir. Two notes on genome rearrangements. *J. Bioinf. Comput. Biol.*, 1(1):71–94, 2003.

[5] G. Tesler. Efficient algorithms for multichromosomal genome rearrangements. *J. Comput. Syst. Sci.*, 65(3):587–609, 2002.

[6] S. Yancopoulos, O. Attie, and R. Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346, 2005.

A Novel Scoring Scheme to Evaluate Match of Peptide and Mass Spectrum

Yantao Qiao, Shiwei Sun, Gongjin Dong, Yu Lin,
Chungong Yu, Dongbo Bu[1]

1 Introduction

Most of the current peptide identification techniques suffer from the inaccuracy in theoretical spectrum prediction for the reason that the fracture mechanism of a peptide is not very clear so far. We presented a model taking the fragment probability of the peptide bond into account [1]. Under this model, we predicted a normalized theoretical intensity vector using our learned parameters. By explaining the peaks generated from the fragment event at specific peptide bond, the experiment spectrum can be transferred into a intensity vector as well. Here, a novel score scheme is used to evaluate the similarity of these two vectors which contain much intensity information and this scheme gives a better result in our experiments compared with other common methods, i.e. the Pearson Correlation Coefficient and Cosine Correlation. An open source package PI (Peptide Identifier) can be downloaded freely from `http://www.bioinfo.org.cn/MSMS`.

2 Event Model with Jensen-Shannon Divergence

In our model, we assume that the fragmentation probability of a peptide bond is decided by the type of specific peptide bond and the position in the peptide. For a peptide $P = X_1X_2 \dots X_L$ with L amino acids, f_i represents the fragmentation reference at the ith position, and the $E(Xaa, Yaa)$ means the fragment probability of the $Xaa - Yaa$ peptide bond. Then, we can derive the theory vector $V^T = \{v_1^T, v_2^T, ..., v_L^T\}$ and the experiment vector $V^E = \{v_1^E, v_2^E, ..., v_L^E\}$, where $v_i^T = \alpha \times f_i \times E(Xaa, Yaa)$ denotes the probability of the fragmentation event at the ith position in a peptide, and v_i^E is the normalized intensity which can be explained in the spectrum by the ith fragmentation event. Here, we solve a non-linear programming problem to train these parameters with a training dataset.

With the learned parameters, we can predict a theory vector V^T for a peptide, and transfer a spectrum to a experiment vector V^E by explaining fragment event. We applied the Jensen-Shannon divergence [3] which can be defined as $JSD = \frac{1}{2}\sum_{i=1}^{L}\left(v_i^T \log \frac{2v_i^T}{v_i^T+v_i^E} + v_i^E \log \frac{2v_i^E}{v_i^T+v_i^E}\right)$ as our scoring function to evaluate the similarity of these two vectors in our experiments.

3 Result of Event Model

We use the LTQ and QSTAR datasets from Gygi to test our scoring function [4]. In each experiment, we take some high reliability matches selected by other software as our training sets, other matches as testing sets. Figure 1 showed that our method can improve the results of the software, such as SEQUEST and Mascot, and other evaluation methods, i.e. Pearson Correlation Coefficient and Cosine Correlation [2].

4 Event Model with EM Algorithm

We use the EM algorithm to learn the neutral loss probabilities of amino acids, and use these probabilities to predict the theoretical spectrum [5]. When comparing an experiment spectrum with a theoretical spectrum, Jensen-Shannon divergence also represents a good performance on the same datasets from Gygi (See Figure 2).

[1]Bioinformatics Lab, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China. Email: {qiaoyantao, dwsun, donggongjin, linyu, cgyu, bdb}@ict.ac.cn

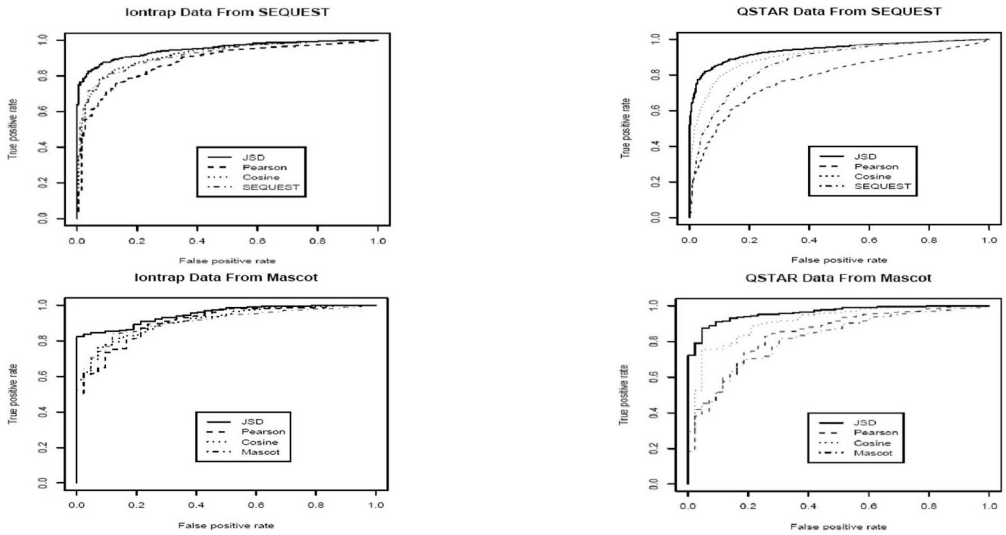


Figure 1: Performance of validating results from SEQUEST and Mascot on LTQ and QSTAR data sets with event model.

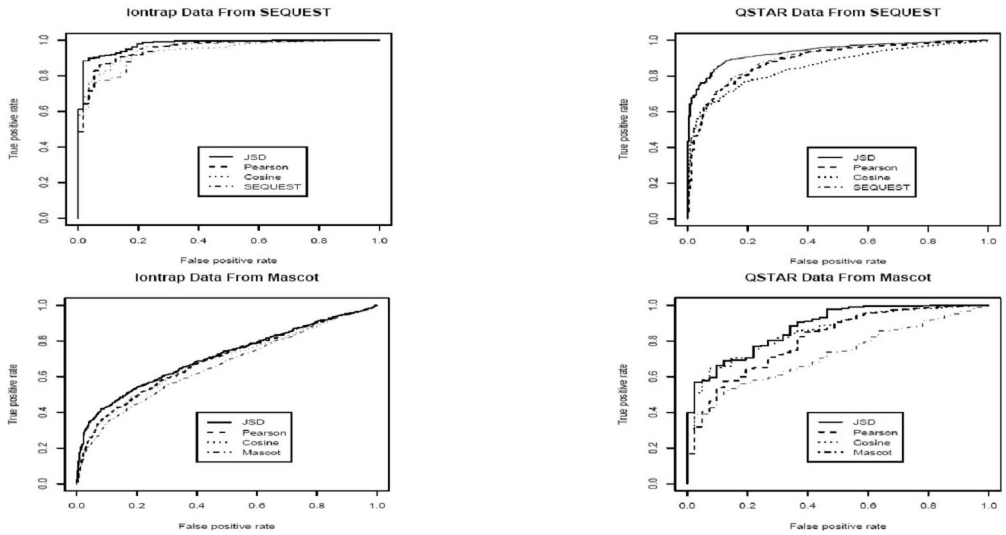


Figure 2: Performance of validating results from SEQUEST and Mascot on LTQ and QSTAR data sets with EM algorithm and event model.

References

[1] Yu CQ, Lin Y, Sun SW, et al. (2007) *J Bioinformatics and Computational Biology*, 5(2), 297–311.

[2] Correlation, `http://en.wikipedia.org/wiki/Correlation`

[3] Jensen-Shannon Divergence, `http://en.wikipedia.org/wiki/Jensen_Shannon_Divergence`

[4] Elias JE, Hass W, Faherty BK, Gygi SP (2005) *Nature Methods*, 2(9), 667–675.

[5] Sun S, Yu C, Qiao Y, et al. (2008) *J. Proteome Res.*, 5(1), 202–208.

Sequence Complexity Measures for Alignment Free Genome Comparisons

Yuriy L. Orlov[1]

1 Applications of DNA Sequence Complexity

The classical problem of biological sequences comparison has traditionally been assessed by pairwise (or multiple) alignment. The task of identifying functional relationships between large, greatly diverged or highly repeated sequences demands "alignment-free" measures of sequence similarity [1, 2] including compression and complexity measures. Several measures of text complexity including combinatorial, linguistic and Lempel-Ziv estimates were implemented [3]. The low complexity may be preconditioned by strong inequality in nucleotide content (biased composition), by tandem or dispersed repeats, by palindrome-hairpin structures, as well as by combination of all these factors. We use the measures developed for mitochondrial genomes analysis, coding/non-coding and regulatory region analysis. Plant mitochondrial genomes have an average a less complexity sequences in comparison with nuclear chromosomes in the same species. As a rule low complexity regions don't contain gene coding regions. A correlation of the nucleosome potential estimate with text complexity was established [4].

2 Algorithms and Statistical Analysis

Analysis of genomic sequences issues the challenge to search for the regions with the low text complexity, which could be functionally important [5]. Low complexity regions are often treated as the regions of biased composition containing simple sequence repeats [6]. The problem of local sequence complexity becomes more significant with developing new genome-wide high-throughout sequencing technologies and whole genome arrays. Process of mapping tags to the reference genome can bias the analysis toward genomic regions with unique and complex sequence patterns, what demand filtration and probabilistic analysis of such regions. Regarding to alignment-free comparisons, widely used approach is based on word frequencies (sequence words, k-tuples). Another set of alignment-free methods is based on information theory approach, the so-called Kolmogorov complexity.

Conditional Kolmogorov complexity $K(X|Y)$ is defined as the length of the shortest program computing sequence X on input Y [7]. Kolmogorov complexity is a non-computable, and in practical applications it is approximated by the length of the compressed sequence calculated by a compression algorithm like LZW [8]. The Lempel-Ziv compression may be interpreted as representation of a text in terms of repeats. Based on this approach, Internet-available tools "Complexity" were developed [3].

The scheme of symbol sequence presentation by Lempel and Ziv was used to measure complexity of sequence by the number of steps of generating process. The permitted operations here are generation of a new symbol (this operation is necessary at least to synthesize the alphabet symbols) and direct copying of a fragment from the already generated part of the text. Copying implies search for a prototype (repeat in a common sense) in the text and extension of the text by attaching the "prepared" block. The scheme for generating the sequence X may be represented as a concatenation H of the fragments:

$$H(X) = S[1 : i_1], S[i_1 + 1 : i_2], \ldots, S[i_{k-1} + 1 : i_k], \ldots, S[i_{m-1} + 1 : N], \quad (1)$$

where $S[i_{k-1} + 1 : i_k]$, is the fragment (component) generated at the kth step (a sequence of elements located from the position $i_{k-1} + 1$ to i_k); N, the length of sequence.

The scheme with minimal number of steps m generating the process should be selected. This scheme determines the complexity of sequence X. The minimal number of components in (1) is provided by selection at each step of the maximally long prototype in the previous history. The complexity decomposition of a sequence is performed from left to right. The algorithm implementation for DNA research was described in details in [3]. Example of complexity analysis for sequence AGAGAGTCCCACATACGAGA is presented in Figure 1.

[1]Genome Institute of Singapore, Singapore. Email: orlovy@gis.a-star.edu.sg

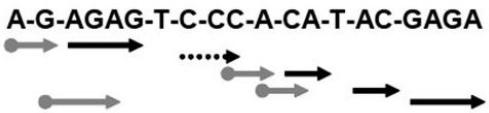


Figure 1: Example of complexity decomposition (grey arrows present prototypes of the fragments with underlying black arrows).

We construct the complexity profile in the sliding window with the length N, the evaluation of complexity is calculated as the whole number $CLZ(X)$ of components of complexity decomposition in the window N, or as the relative number of the components $CLZ(X)/N$. For the example in Figure 1, normalized complexity value $CLZ(X)/N = 11/20 = 0.55$. Length N may vary from tens nucleotides to megabases. Other complexity measures such as entropy and linguistic complexity correlate [3] and could be used mainly by sliding window measure. Complete sequence decomposition allow reveal and non-redundantly present patterns of sequence/chromosome structure.

3 Discussion

Results on complexity measure application refer to different domains: from genome structure investigation to extended regulatory region comparison. Recently compositional complexity measures were applied to periodicity patterns [9]. Alignment-free measures might be useful as pre-selection filters for alignment-based querying in large-scale applications.

Comparison and complexity analysis of plant mitochondrial (mt) genomes revealed mosaic structure. In mitochondrial genome of *A.thaliana* sequences of nuclear origin represent 4% of the sequence. About 2% of the genome is composed of unaccounted sequences found both in the nucleus and the mitochondria and 1.2% of the genomic DNA are sequences of plastid origin. We found such regions by compositional bias. Low complexity regions in *Beta vulgaris* (sugar beet) mt genome contain tRNA genes. Other mt genomes also shown similar distribution of large low complexity regions.

Overall our analysis of genomic DNA shown high correlation of Lempel-Ziv and linguistic measures of text complexity. Entropy measures have less correlation with Lempel-Ziv estimation for large sliding window size ($>$ 1Kb). At average mt genomes have less complexity by Lempel-Ziv measure than such complexity estimations for chromosomal sequences. This observation suggests greater presence of long repeats in mitochondrial genomes. Note, for example, extra large size of maximal exact repeat 43.7Kb in rice mt genome. This value is unique for sequences of such size. (Another two bacterial genomes have exact repeats greater than 40Kb *E.coli* (strain O157 H7) and *S.agalactiae*, strain NEM316).

References

[1] Kantorovitz, M.R., Robinson, G,E. and Sinha, S. 2007. A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, 23(13):i249–i255.

[2] Höhl, M. and Ragan, M.A. 2007. Is multiple-sequence alignment required for accurate inference of phylogeny? *Syst Biol*, 56:206–221.

[3] Orlov, Y.L. and Potapov, V.N. 2004. Complexity: An internet resource for analysis of DNA sequence complexity. *Nucleic Acids Res*, 32:W628–W633.

[4] Orlov, Iu.L., Levitskii, V.G., Smirnova O.G. et al. 2006. Statistical analysis of nucleosome formation sites. *Biofizika*, 51:608–614.

[5] Hancock, J.M. 2002. Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. *Genetica*, 115:93–103.

[6] Orlov, Y.L., Te Boekhorst, R., Abnizova, I.I. 2006. Statistical measures of the structure of genomic sequences: Entropy, complexity, and position information. *J Bioinform Comput Biol*, 4:523–536.

[7] Li, M. and Vitányi, P.M.B. 1997. *An Introduction to Kolmogorov complexity and its Applications*, NY: Springer Verlag.

[8] Lempel, A. and Ziv, J. 1976. On the complexity of finite sequences. *IEEE Trans Inf Theory*, 22:75–81.

[9] Bolshoy A. 2008. Revisiting the relationship between compositional sequence complexity and periodicity. *Comput Biol Chem*, 32:17–28.

A Gene Ontology-Based Method to Present Pathway Relations

Chia-Lang Hsu,[1] Yen-Hua Huang,[2] Ueng-Cheng Yang[1,2,3,*]

1 Introduction

Biology researches usually start with identifying a gene set that relates to a given phenomenon. The order of interactions between these gene products is a pathway. Those pathways that share components imply cross talks. These related pathways are the basis for pathway integration, which is essential for reconstructing a regulatory network. Since different pathways may contain different numbers of genes, the absolute number of shared component is not a good way to present relation. To address this issue, we propose a gene ontology-based method that aims to evaluate whether any pair of pathways is related functionally.

2 Rationale

A vector space model has been used to compute similarity between pairs of pathways [1]. Since a pathway is a gene set, we have extended this vector space model to determine the relatedness between pairs of pathways. A pathway is represented by a specific vector p_j as follows:

$$p_j = (w_{1,j}, w_{2,j}, \ldots, w_{n,j})$$

where $w_{i,j}$ is the weighting value that the GO term i takes on for pathway j. We only consider the biological process ontology to construct the vector. Since each GO term may annotate more than one gene in a pathway, we need to consider both the term frequency (tf) and inverse document frequency (idf). The weighting value of each GO term i is thus defined by using tf * idf method:

$$w_{i,j} = tf_{i,j} * idf_i = \frac{freq_{i,j}}{n_j} * \log \frac{N}{n_i}$$

where $freq_{i,j}$ is the number of components in the pathway j annotated by the term i and the children term of term i, n_j is the number of components in the pathway j, n_i is the number of proteins annotated by term i in human, and the N is the total number of proteins in human. The functional relatedness between two pathways, $sim(p_1, p_2)$, is defined as follows:

$$sim(p_1, p_2) = \frac{\vec{p_1}\vec{p_2}}{|p_1||p_2|}$$

The functional relatedness is computed pairwise for all the pathways from BioCarta [4]. The GO term annotation of each component of pathways is derived from NCBI Entrez Gene. We used a set of genes published by West et al [2] to test our proposed method.

3 Results and Discussion

We obtained 314 pathways from BioCarta and generated 98,282 pairs of pathways. Only 8,770 pairs contain one or more components between two pathways. The Pearson's correlation coefficient of shared components and the functional relatedness is only 0.63 and it means functional relatedness is not strongly correlated with shared components of two pathways. In order to determine the threshold, we randomly generated gene sets with size from 10 to 50 and computed the functional relatedness with BioCarta

[1]Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan.
[2]Institute of Biochemistry and Molecular Biology, National Yang-Ming University, Taipei, Taiwan.
[3]Center for Systems and Synthetic Biology, National Yang-Ming University, Taipei, Taiwan.
* Corresponding author: yang@ym.edu.tw

pathways. The mean of relatedness of random set is 0.527 and the maximum of that is 0.740, so we set threshold as 0.7. The gene set published by West et al [2] was used to draw a relation graph. This set contains 94 genes that showed significant difference in gene expression level between two breast cancer subtypes—the presence or absence of estrogen receptor (ER). Cytoscape [3] was used to visualize the relation of pathways. As shown in Figure 1a, this relation graph contains 50 pathways and 458 relations, i.e. the two pathways share one or more components. This pathway relation graph is too complex to interpret. If, however, our gene ontology-based method was used (Figure 1b), the noise is reduced and it is more clear for us to understand the relation the pairs of pathways.

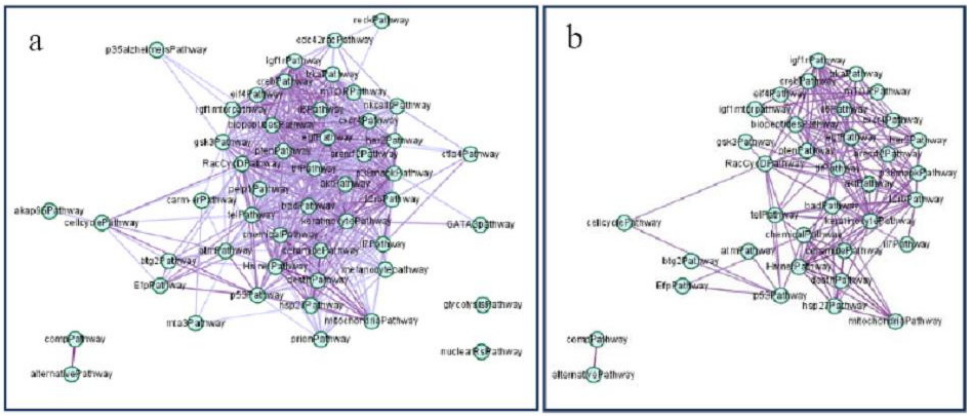


Figure 1: Pathway relation graph. Nodes indicate the pathways and edges indicate relation between two pathways. a) relations found by shared components; b) relations found by gene ontology-based methos.

4 Conclusions

This approach greatly simplifies the relation among pathways presented by shared components, but preserve the biologically significant relations. Therefore, this method can be used to display the relation among a group of pathways, which are extracted by using state-specific gene expression information. The ER+ and ER- states of breast cancer have been used as an example to demonstrate the usefulness of this method.

References

[1] Chabalier, J., Mosser, J., and Burgun, A., 2007. A transversal approach to predict gene product networks from ontology-based similarity. *BMC Bioinformatics*, 8:235

[2] West, M., Blanchette, C., Dressman, et al. 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci.*, 98:11462–11467.

[3] Shannon, P., Markiel, A., Ozier, O., et al. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498–2504.

[4] BioCarta, `http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways`

Information-Guided Knowledge-Based Potential Functions for Protein Structure Prediction

Armando D. Solis,[1] S. Rackovsky[2]

1 Introduction

Success in the computational protein structure prediction relies on score functions or potentials to evaluate the native fitness of a given sequence-structure alignment. Among the myriad approaches, "knowledgebased" potentials, referring to a class of methods that utilize experimentally derived sequence and structure data to derive function parameters, have gained prominence in large part because of their reduced computational complexity, as compared to more *ab initio* methods. Knowledge-based potentials succeed primarily by simplifying the description of sequence and structure, as well as their interactions, making it readily derived from data and applied in a diversity of prediction efforts. Not surprisingly, the ad-hoc nature of knowledge-based methods has resulted in an overwhelming diversity of potential functions that span the range of theoretical rigor, detail, and performance. Faced with a finite set of structural data (in the PDB) from which to derive these functions, optimization of data use becomes a worthy goal.

The principal objective of our recent and current work [1-4] is to establish a systematic and rigorous method for constructing the best performing knowledge-based potentials, in light of the pressure from limited data. These potential functions, whether proper "energy" functions or probabilistic quantities, are fundamentally information-theoretic functions, a correspondence we have demonstrated rigorously [1, 2]. By recognizing that deriving potential functions from prior data is primarily an informatic concern, any physical justification required by the "energetic" viewpoint is conveniently bypassed, replaced instead by a greater flexibility to choose the best functional form and parameterization. Once the problem is recast in an information-theoretic framework, it becomes straightforward to design potential functions by direct and automatic maximization of the information which can be extracted from existing data. The key point, as we have firmly established [1, 2], is that knowledge-based potentials optimized for information extraction show increased performance in comprehensive fold recognition/threading tests.

2 The Complete Information Equation

Our primary task in developing a better strategy to formulate knowledge-based potential functions is to examine how these score functions relate to the basic concept of information storage in protein sequences. The powerful maxim, that all the information needed to specify the native conformation (C) of a globular protein resides in its amino acid sequence (S), can be translated into an information-theoretic framework. Structural information encoded in sequence can be expressed using mutual information between S and C:

$$I(C,S) = \sum_{(C,s)} p(C,S) \ln \frac{p(C,S)}{p(C)p(S)} \approx \frac{1}{n_S} \sum_{\text{all seq}} \ln \frac{p(C|S)}{p(C)} \tag{1}$$

The right-most quantity is an empirical estimate of mutual information, formed from an average over all (C,S) pairs in the protein universe, or more specifically, from a representative set of ns natural sequences in the data set.

If the conformation can be described by a set of desciptors $C = \{c_1,\ c_2,\ ...,\ c_t\}$ (a combination of structural states like side chain orientation, backbone trace, contact map, etc.) and the sequence as a chain of letters $S = \{s_1,\ s_2,\ ...,\ s_w\}$, then the equation above expands to what we shall henceforth call the "complete information equation" of protein folding:

$$I(C,S) = \frac{1}{n_S} \sum_{\text{all seq}} \left(\ln \frac{p(c_1|s_1,s_2,\ldots,s_w)}{p(c_1)} + \sum_{i=2}^{t} \ln \frac{p(c_i|s_1,s_2,\ldots,s_w,c_1,c_2,\ldots,c_{i-1})}{p(c_i|c_1,c_2,\ldots,c_{i-1})} \right) \tag{2}$$

[1]Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, New York, NY, USA. Email: armando.solis@mssm.edu

[2]Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, New York, NY, USA. Email: shelly@camelot.mssm.edu

which contains the fullest expression of sequence-dependent conformational propensities. It is easy to see that each term of the inner summation, including the first term, is the complete form of the score functions commonly used in protein structure prediction.

A comparison with traditional knowledge-based potentials allows us to understand the kinds of simplifications made (unwittingly) by those models which partition the potential into independent interactions. Foremost, the effect of the *entire length* of the sequence is always truncated in the conditional probability to include only the most dominant subsequences. While limitations of the size of the database is a crucial consideration in restricting sequence description, this guiding equation argues that more effort should be expended if one wants to develop the most accurate of knowledge-based potentials. An equally key issue is the common practice of excluding the effect of other structural features in both the conditional and the reference probabilities. Sequence is used exclusively to condition the structural propensities in these score functions, but the complete information equation demonstrates that correlative effects of other structural features are a factor in scoring. Structural correlations can take on two types: those that occur within the same descriptor (in this case, the successive phi-psi dihedral angles in the backbone chain) or those that occur across different descriptors (e.g., residues far in sequence but make close contact with the residue in question in the folded environment). While these phenomena have been recognized by some and incorporated in a number of score functions, there remains an opportunity to organize and systematize of these correlative structural factors.

3 Current Developments and Future Direction

We are currently developing a comprehensive informatic strategy, and apply it to a number of current relevant problems in protein structure prediction. Our over-all goal is to derive the most informative total potentials, something that has not been done before. We envision the important features of these potentials to be: (a) they will be exhaustive, aiming to include *all* local and non-local sequence and structure descriptions and their correlations; (b) they will automatically reduce redundancies between overlapping descriptions (e.g., between phi-psi dihedral angle pair and reduced a-carbon backbone descriptions), while still incorporating information from all of them; (c) they will be derived from all structural data in the current Protein Data Bank (PDB), adjusting descriptions and parameterizations automatically to respond to the pressures of limited but growing data; (d) they will maximize information retention overall, assuring their superior performance in any type of protein folding application.

In deriving potentials from informatic principles, we are also addressing the reference state problem directly. Deriving the correct reference state, a fundamental problem in energy-based (Boltzmann) approaches, becomes forthright if the problem is restated in terms of optimizing for information retention. The functional form of the reference state that maximizes information gain shall be chosen by our methodology, irrespective of whether it conforms to popular pseudo-theoretical justifications.

As a next step, we are also designing the next generation of knowledge-based potentials: true query-specific potentials, whose parameters will be tailored automatically for any sequence of interest. We can demonstrate, in conjunction with the reference state study and the maximum information approach, the significant advantage of potentially including the contribution of every structure in the PDB to the resulting potential, weighted with respect to their relative similarity to the query sequence.

References

[1] Solis, A. and Rackovsky, S. 2008. Information and Discrimination in Contact Potentials. *Proteins: Structure, Function, and Bioinformatics*, in press (published online, November 14, 2007).

[2] Solis, A. and Rackovsky, S. 2006. Improvement of Statistical Potentials and Threading Score Functions Using Information Maximization. *Proteins: Structure, Function, and Bioinformatics*, 62:892–908.

[3] Solis, A. and Rackovsky, S. 2002. Optimally Informative Backbone Structural Propensities in Proteins. *Proteins: Structure, Function, and Bioinformatics*, 48:463–486.

[4] Solis, A. and Rackovsky, S. 2000. Optimized Representations and Maximal Information in Proteins. *Proteins: Structure, Function, and Bioinformatics*, 38:149–164.

Extraction and Grounding of Protein Mutations via Ontology-centric Knowledge Integration

Rajaraman Kanagasabai,[1] Christopher Baker[2]

1 Introduction

Rich information on point mutations is scattered across heterogeneous data resources. In order to convert this information into actionable knowledge, the researcher has to extract this information and map the mutation to its associated protein. This task is complicated by the heterogeneous document formats, languages style and slow adoption of common and standardized vocabulary to describe the mutations. Consequently integrating this information is time-consuming and manually curated databases have been shown to contain errors in the order of 40% [4]. A number of mutation extraction systems have been developed to automate the retrieval of this information [1, 3]. In some cases the specific goal is the mapping of mutations to proteins for visualization in 3D [2, 5] requiring accurate identification and retrieval of sequences from protein databases. Here we propose an improved algorithm for extraction and mapping of mutations in an ontology-centric framework [6], and examine the efficiency of targeted sequence retrieval for proteins reported as mutated.

2 Methodology

In [5] we described a multi-tier system to automate the workflow required to support extraction mutation and structure annotation with mutation mentions. The system coordinates two main pipelines: (i) the ontology population workflow comprising of document retrieval, information extraction, data integration, and ontology instantiation (ii) the ontology employment workflow comprising; query of the populated ontology, protein structure coordinate retrieval, homology modeling if a structure is unavailable, mutant residue mapping and protein structure visualization. A key subsystem in the ontology population workflow instantiates protein names and point mutations extracted through text mining and mines the ontology instances based on a cross validation between protein sequences retrieved from an NCBI protein name search and the successful mapping of text-derived mutations onto the sequences (See Figure 1).

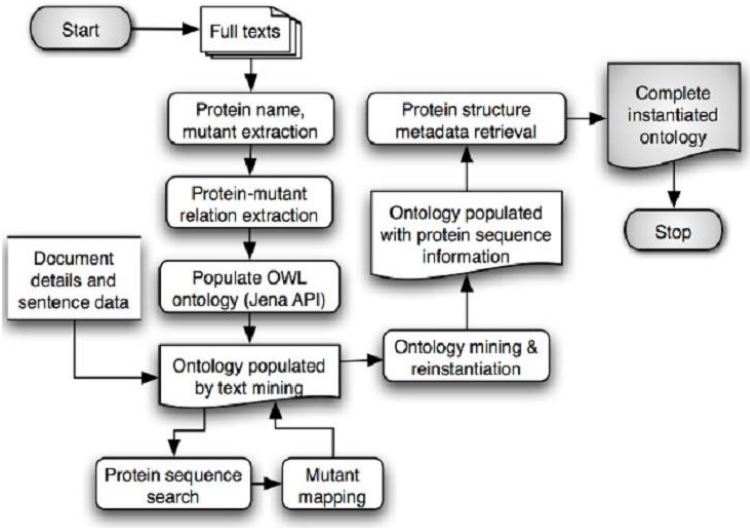


Figure 1: Mutation extraction and mapping workflow.

[1]Data Mining Department, Institute for Infocomm Research, Singapore. Email: kanagasa@i2r.a-star.edu.sg
[2]Data Mining Department, Institute for Infocomm Research, Singapore. Email: cbaker@i2r.a-star.edu.sg

Following the same framework, here we propose an improved algorithm for the extraction and mapping of mutations by integrating additional domain knowledge. The improved strategy employs: 1) a more comprehensive protein name dictionary compiled from a custom-cleaned version of Uniprot, 2) a better protein name normalization technique to resolve synonyms and abbreviations, and 3) the extraction of organism names from the texts using a list of NCBI Taxonomy terms as the dictionary, and use of the dictionary to filter out false positives in the protein sequences retrieved. Furthermore, we employ additional heuristics (such as using task-specific keywords to expand the protein search query) to improve the precision. To benchmark the performance of our system we used the Protein Mutant Database (PMD[3]) as a gold standard. We collected PMD records reporting two mutated protein families namely, phosphatases (47 records) and kinases (45 records), that had at least 2 point mutations and a MEDLINE citation whose full text paper was download-able with our content acquisition engine. With this corpus, we evaluated two tasks: 1) Protein-Mutant Tuple Extraction and 2) Mutation Grounding, and measured performance by computing task-specific precision and recall [7]. The results are presented in Table 1.

Task	Phosphatases		Kinases	
	Precision	Recall	Precision	Recall
Protein-Mutant Tuple Extraction	81%	74%	73%	58%
Mutation Grounding	78%	71%	67%	46%

Table 1: Performance evaluation.

We observed marginal improvements in the performance ($< 3\%$) for phosphatases. However, in the cases of kinases, our earlier algorithm resulted in a 32% recall for Task 1 and 22% recall for Task 2. By effectively handling the term ambiguities in kinase protein names and incorporating additional domain knowledge, the new algorithm achieved significant improvements in the performance.

3 Discussion

Our work addresses a pivotal step in the workflow facilitating the provision of mutation mentions as protein structure annotations. In particular we have employed a diversity of approaches to improve our algorithm and tested it for the transfer of mutation mentions from commercially relevant protein families, albeit with moderate performance in the case of kinases. Our ongoing work involves further improvements and testing the effectiveness of the algorithm for mutation grounding, i.e. mapping mutations and retrieving the correct protein sequences using an offset analysis. Specifically we will examine performance on a wider range of protein families.

References

[1] C.J.O. Baker (Ed). 2007. Special Issue: Making Sense of Mutations Requires Knowledge Management. *Journal of Bioinformatics and Computational Biology*, 5(6).

[2] C.J.O Baker, R. Witte, 2006. Mutation Mining: A Prospector's Tale. *Information Systems Frontiers*, 8(1):47–57.

[3] J. G. Caporaso, W. A. Baumgartner, Jr., D. A. Randolph, et al. 2007. MutationFinder: A high-performance system for extracting point mutation mentions from text. *Bioinformatics*, 23:1862–1865.

[4] J. G. Caporaso, N. Deshpande, J. L. Fink, et al. 2008. Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. *Pacific Symposium on Biocomputing*, 13:640–651.

[5] R. Kanagasabai, K.H. Choo, S. Ranganathan, C.J.O. Baker. 2007. A workflow for mutation extraction and structure annotation. *Journal of Bioinformatics and Computational Biology*, 5(6):1319–1337.

[6] R. Witte, T. Kappler and C.J.O. Baker. 2007. Enhanced semantic access to the protein engineering literature using ontologies populated by text mining. *International Journal of Bioinformatics Research and Applications*, 3(3):389–413.

[7] R. Witte and C.J.O. Baker. 2007. Towards a systematic evaluation of protein mutation extraction systems. *Journal of Bioinformatics and Computational Biology*, 5(6):1339–1359.

[3] http://pmd.ddbj.nig.ac.jp

Improving Detection Performance for Gene Set Functional Enrichment and Finding Transcription Factor Binding Sites

Przemysław Biecek,[1] Adam Zagdański,[2] Rafał Kustra,[3] Stanisław Cebrat[4]

1 Introduction and Methodology

We propose a novel method for adjusting p-values when probability of being false for each hypothesis is a priori known or can be estimated. We show theoretical basis of our method. Two applications in gene set functional enrichment and in finding transcription factor binding sites are presented. Results show that our method significantly improves predictive performance and decreases false positive rate.

Let us consider a set of m null hypotheses $H_0^{(i)}$. Let $\pi_1^{(i)}$ denote the known probability that $H_0^{(i)}$ is false, while $p^{(i)}$ stands for p-value corresponding to this hypothesis. Using the Bayes theorem we get

$$Pr(H_i = 0|p^{(i)}) = \frac{p^{(i)}(1-\pi_1^{(i)})}{p^{(i)}(1-\pi_1^{(i)}) + F_i(p^{(i)})\pi_1^{(i)}} \leq p^{(i)} \frac{(1-\pi_1^{(i)})}{R_i(p^{(i)})\pi_1^{(i)}} = \frac{p^{(i)}}{OR^{(i)}F_i(p^{(i)})}$$

where $OR^{(i)} = \pi_1^{(i)}/(1-\pi_1^{(i)})$ (called "odds ratio") represents how many times the null hypothesis is less probable than the alternative while $F_i(p^{(i)})$ corresponds to p-value distribution for true ith alternative. We propose to use the quotient of p-value and OR (as a reasonable approximation of $Pr(H_i = 0|p^{(i)})$ instead of the original p-values. Term $F_i(p^{(i)})$ is neglected since it cannot be directly estimated. The advantage of such approach lies in incorporating the known or estimated false null frequencies. These weights may be normalized (see [1], Theorem 1) in a way which allows to control FDR. Finally, we propose to compute the adjusted p-values in the following way

$$p_{adj}^{(i)} = p_{org}^{(i)} \frac{\sum_{j=1}^{m} OR^{(j)}}{mOR^{(i)}}. \tag{1}$$

Note, that ordering of adjusted p-values may be different from the ordering of original ones. Below, using two examples, we show that the adjusted p-values give better results.

2 Applications and Results

Gene Set Functional Enrichment Analysis (GSFEA). The main purpose of the GSFEA is to identify the biological attributes, shared by a given set of genes (query set), that distinguish them from the remaining ones (reference set). A set of genes is deemed to be functionally enriched for a given attribute if the fraction of genes within this set sharing that attribute is larger than could be expected by chance. The enrichment is typically ascertained using the one-tailed Fisher's exact test (for more details see e.g. [2]). Enrichment p-values are used to rank functional categories predicted for a gene.

Table 1 presents 10 first level GO terms in Biological Process ontology. For each term one may estimate frequencies of false nulls and corresponding odds ratios (from annotation databases). We apply GSFEA to these GO terms and set of 1730 genes derived from yeast protein-protein interactions dataset (eg. from BioGRID database). Figure 1 shows summary of outcomes from original p-values and adjusted p-values computed from equation (1). Using different error measures (ROC, Precision, Recall) we show that results obtained for the modified p-values outperform results for the original p-values.

Finding Transcription Factor Binding Sites (TFBS). An important research is the identification of transcription factor binding sites. The main step here is to test (eg. using likelihood ratio test) that a given DNA sequence is significantly related to a given motif. Many different tests are used in such case (see e.g. [3]). They incorporate information specific to given motifs and sequence (eg. length, nucleotide composition etc.) but do not incorporate known or estimated frequencies of motifs.

[1]Department of Genomics, Wrocław University, Poland. Institute of Mathematics of the Polish Academy of Science, Poland. Email: `przemyslaw.biecek@gmail.com`

[2]Institute of Mathematics and Computer Science, Wrocław University of Technology, Poland. Email: `zagdan@pwr.wroc.pl`

[3]Department of Public Health Sciences, University of Toronto, Canada. Email: `r.kustra@utoronto.ca`

[4]Department of Genomics, Wrocław University, Poland. Email: `cebrat@smorfland.uni.wroc.pl`

From database of known TFBS [4] one may estimate these frequencies. As mentioned, they differ markedly among motifs (see example motifs frequencies in Table 2). In Figure 2 (left) we compare FDR for different motifs as function of the threshold α. For given α average FDR is different for motifs with different OR. Figure 2 (right) shows the average FDR computed for all considered motifs. If we neglect information about differences in frequencies (original p-values), the rate of false signals is higher than for adjusted p-values.

Conclusion. The fraction of false positives may be significantly reduced by using adjusted p-values.

References

[1] Ch. Genovese, K. Roeder, L. Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

[2] P. Khatri, S. Drăghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, 2005.

[3] W. Krivan, W.W. Wasserman. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Research*, 11(9):1559–1566, 2001.

[4] E. Blanco, D. Farre, M. Alba, X. Messeguer, R. Guigo. ABS: A database of Annotated regulatory Binding Sites from orthologous promoters. *Nucleic Acids Research*, 34:D63–D67, 2006.

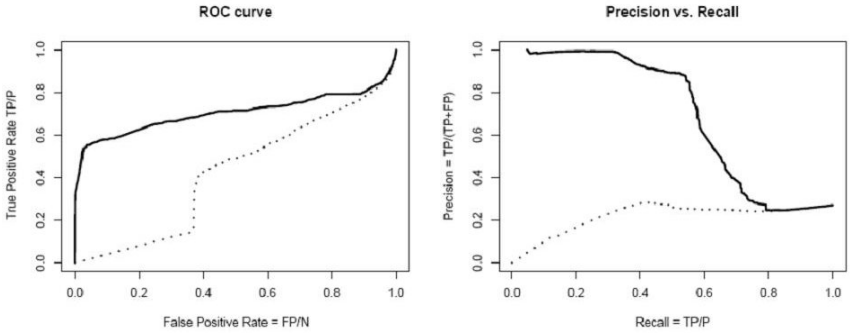


Figure 1: Summary of GSFEA outcomes. Dotted line corresponds to results for original p-values, solid line for adjusted p-values.

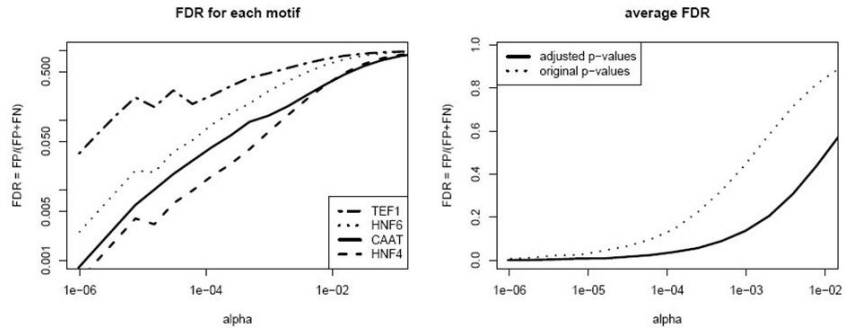


Figure 2: Summary of TFBS outcomes. Dotted line corresponds to results for original p-values, solid line for adjusted p-values.

GO term	π_1	OR	function	GO term	π_1	OR	function
GO:0000003	0.06637	0.07109	reproduction	GO:0050896	0.13085	0.15054	response to stimulus
GO:0022610	0.00402	0.00403	biological adhesion	GO:0040011	0.03094	0.03193	growth
GO:0032501	0.09495	0.10491	multicellular organismal process	GO:0032502	0.10510	0.11745	developmental process
GO:0009987	0.92726	12.7468	cellular process	GO:0008152	0.68399	2.16442	metabolic process
GO:0051704	0.02881	0.02967	multi-organism process	GO:0048511	0.00024	0.00024	rhythmic process

Table 1: Estimated frequency of annotations π_1 for ten first-level GO-BP terms.

Motif	# obs	π_1	OR	Motif	# obs	π_1	OR	Motif	# obs	π_1	OR
CAAT	13	0.0200	0.0204	CEBP	37	0.0569	0.0604	TBP	95	0.1462	0.1712
SP1	89	0.1369	0.1586	HNF1	27	0.0415	0.0433	HNF4	10	0.0154	0.0156
HNF6	4	0.0062	0.0062	IRF	2	0.0031	0.0031	TEF1	7	0.0108	0.0109

Table 2: Estimated frequency of specific transcription factors in promoters. Data from ABS database [4].

Structural Strand Asymmetry for Transcription Orientation Prediction in Unaligned ncRNA Sequences

Brian J. Parker,[1] Jiayu Wen,[2] Georg F. Weiller[3]

1 Introduction

Many RNA functions are determined by their specific secondary and tertiary structures. In our previous work [1] we introduced the concept of structural strand asymmetry between complementary strands in ncRNA sequences due to G::U non-canonical base pairings. RNA structures are folded by the canonical G::C and A::U base pairings as well as by the non-canonical G::U complementary bases. G::U base pairings in RNA secondary structures may induce structural asymmetries between the transcribed and non-transcribed strands in their corresponding DNA sequences. This is likely so because the corresponding C::A nucleotides of the complementary strand do not pair. As a consequence, the secondary structures that form from a genomic sequence depend on the strand transcribed, and this information can be used to identify the transcribed strand. In our previous work, we compared both global and local structural formation asymmetry and analyzed it on non-protein-coding transcripts. We investigated this idea further to show that both thermodynamic stability of global RNA structures in the transcribed strand and RNA structure strand asymmetry are statistically stronger than that in randomized versions preserving the same di-nucleotide base composition and length [2].

Such structural strand asymmetry features have potential application in detecting transcribed region and de novo ncRNA prediction in genomic data, and have the advantage that they also can be used to identify the transcription orientation. Other recent approaches to this problem have incorporated the structural strand asymmetry feature in transcribed strand identification in conserved RNA secondary structures using multiple sequence alignments [3]. Yet, the requirement for the presence of multiple sequence alignments is not suitable for detecting transcribed strands in all cases such as non-conserved ncRNAs and where multiple sequence alignment data is otherwise not available.

In this paper, we further characterize the strand asymmetry features described in [1] and analyze their application to transcribed strand detection in sequences without multiple alignment information. Using a machine learning approach, we show that measures of local structural strand asymmetry in combination with base composition asymmetry features can be used to predict the transcription orientation across all the studied non-coding RNA families with classification accuracy approaching 90%.

2 Results and Discussion

RNA intrinsic structural constraints may affect base compositions and cause deviation from approximately A%=T% and C%=G%. Both our previous work [1] and other studies [4] have suggested that base compositional biases may serve as indicators of ncRNAs and transcription. We tested whether a combination of RNA structural strand asymmetry and base compositional asymmetry leads to better classification accuracy for predicting strand direction. Five features representing local RNA structure strand asymmetry and local base composition strand asymmetry were input to a random forests classifier.

In the potential application of the asymmetry feature to genome-wide studies, it is of interest to examine the performance using a fixed-size sliding window, as would be required in this application. We estimate local structural strand asymmetry using RNALfold [5] (in Vienna RNA package) which computes all possible local structures smaller than a fixed window size L. RNALfold was run on the two complementary strands to estimate the structural asymmetry. The mean minimum free energy (MFE) over all the local structures, normalized by the length of the local structures, was computed for each strand and the difference between the two strands was defined as $\Delta\text{MFED}_{tr-ntr}$. $\Delta\text{MFED}_{tr-ntr}$ was calculated on the original RNA sequences and on di-nucleotide shuffled sequences. To get a measure of the strength of the structural asymmetry compared with the dinucleotide shuffled randomized background,

[1]Life Sciences Group, NICTA, University of Melbourne, Australia. Email: brian.parker.phd@gmail.com
[2]Bioinformatics Center, University of Copenhagen, Denmark. Email: jeanwen@binf.ku.dk
[3]Bioinformatics Laboratory, RSBS, Australian National University, Canberra, Australia. Email: georg.weiller@anu.edu.au

we computed the Z-score for $\Delta\text{MFED}_{tr-ntr}$. Z-score $= (x-\mu)/\sigma$, where x is the $\Delta\text{MFED}_{tr-ntr}$ values from the original sequences, μ is the mean value over the di-nucleotide shuffled sequences, and σ is the standard deviation over the shuffled sequences.

Local base composition asymmetry in the predicted structures was estimated using the following features: $(G-C) = (f_G - f_C)/(f_G + f_C)$, $(A-U) = (f_A - f_U)/(f_A + f_U)$, and $\log_2\left(\frac{f_G+f_U}{f_A+f_C}\right)$. Note that these features are standardized such that a value of 0 indicates no asymmetry across strands.

The classiffication discriminative power of the combined features was evaluated using classification accuracy and the area under the receiver operating characteristic (ROC) curve (AUC). The ncRNA sequences used in this study were obtained from the Rfam database (release 8.0). The results of the combination of all features (see Table 1) shows that most ncRNA sequences are classified correctly with an improved classification accuracy of 89% and an AUC of 94% for detecting strand direction. Except for 7SK, all other individual ncRNA families show substantially improved accuracy rates, and particularly miRNAs (93%), 5.8S rRNA (95%), Hammerhead 3 (98%), intron gpI (97%), intron gpII (99%), IRES (91%), Nuclear RNase P (97%), and SRP euk arch (97%). The strand asymmetry feature combination therefore provides a reliable prediction for detecting transcription orientation of unknown ncRNA families in practice. An estimation of feature importance (as measured by the mean decrease in Gini index) in the random forest gave the ranking of features, in decreasing order: local (G-C), $\Delta\text{MFED}_{tr-ntr}$, Z-score of $\Delta\text{MFED}_{tr-ntr}$, $\log_2\left(\frac{f_G+f_U}{f_A+f_C}\right)$, and $(A-U)$.

By comparison, the approach using multiple sequence alignment and structure conservation features as well as structural strand asymmetry in [3], gave an overall accuracy for ncRNA (alignment) classification of 82.2%. Although the test sets used are not directly comparable, the results shown here for the combination of the structural strand asymmetry and base asymmetry features shows approximately similar if not higher performance.

References

[1] Wen, J., Parker, B. J. and Weiller, G. F. 2007. *In silico* identification and characterization of mRNA-like noncoding transcripts in *Medicago truncatula*. *In Silico Biology*, 7, 0034.

[2] Wen, J., Parker, B. J. and Weiller, G. F. 2008. Analysis of structural strand asymmetry in non-coding RNAs. In: *Proc. 6th Asia-Pacific Bioinformatics Conference* (APBC 08), pp. 187–198.

[3] Reiche, K. and Stadler P. F. 2007. RNAstrand: Reading direction of structured RNAs in multiple sequence alignments. *Algorithms for Molecular Biology*, 2:6.

[4] Glusman, G., Qin, S., El-Gewely, M. R., et al. 2006. A third approach to gene prediction suggests thousands of additional human transcribed regions. *PLOS Computational Biology*, 2(3):160–173.

[5] Hofacker, I. L., Priwitzer, B., and Stadler, P.F. 2004. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, 20:186–190.

ncRNA type	Predictive accuracy using $\Delta\text{MFED}_{tr-ntr}$			Classification using feature combination	
	No. seq	original	di-shuffled	Accuracy	AUC
ncRNA	10423	78%	67%*	89%	0.94
miRNA	1000	85%	77%*	93%	0.99
5.8S rRNA	1000	89%	68%*	95%	0.98
5S rRNA	1000	81%	64%*	89%	0.96
7SK	171	64%	48%*	49%	0.49
Hammerhead 1	62	48%	34%*	55%	0.62
Hammerhead 3	252	65%	57%*	98%	0.99
Intron gpI	1000	76%	55%*	97%	0.99
Intron gpII	1000	98%	77%*	99%	1.00
IRES	1000	77%	69%*	91%	0.97
RNase MRP	45	64%	52%*	87%	0.89
Nuclear RNase P	110	71%	68%	97%	1.00
snoRNA CD-box	1000	58%	60%	77%	0.84
snoRNA HACA-box	1000	78%	68%*	77%	0.86
SRP euk arch	430	87%	88%	97%	1.00
tmRNA	350	73%	59%*	84%	0.93
tRNA	1000	72%	67%*	83%	0.89
mRNA	614	65%	51%*	72%	0.80

Table 1: *Statistically significant $\Delta\text{MFED}_{tr-ntr}$ difference between original and di-nucleotide shuffled.

Zero Recombinant Haplotype Inference with Missing Data

Xin Li,[1] Jing Li[1]

1 Introduction

Haplotype information is important in representing human genetic variation and in association mapping of complex traits. However, humans are diploid and in practice, genotype data instead of haplotype data are collected directly. Therefore efficient and accurate computational methods for haplotype reconstruction are needed and have been investigated intensively recently. We study the problem of zero recombinant haplotype configuration from pedigree data. We formulate the problem as a linear system of inheritance variables and use disjoint-set structures to encode the connectivity information and to detect constraints from the pedigree. Another disjoint-set structure is used to encode and check the consistency of constraints. By doing so, our algorithm can output a general solution in near linear time $O(mn\alpha(n))$ on a tree pedigree without missing data, where m is the number of loci, n is the number of individuals and α is the inverse Ackermann function. This is a further improvement over existing ones. For a looped pedigree or a pedigree with missing data, we can directly extend the above approach by considering existing constraints on inheritance variables. The search space thus is dramatically reduced. The algorithm has been implemented into a software program and experiment results show it can effectively infer all haplotype solutions for a pedigree with 128 members over 200 loci with 20% missing within 0.1 second. Comparisons with other two popular programs show that it achieves 10–10^5 fold improvements over different settings. The experimental study also provides some empirical evidence on the complexity bounds suggested by theoretical analysis. The program is available at `http://www.eecs.case.edu/jxl175`.

2 Methods

Definition 7 *ps variable $p_i^x \in \{0,1\}$ is defined for each locus i of each individual x. $p_i^x = 0$ if the smaller allele of locus i is of paternal source, $p_i^x = 1$ if it is of maternal source. We technically let $p_i^x = 0$ if locus i is homozygous (two alleles being the same).*

Definition 8 *Inheritance variable $h^{x_1x_2} \in \{0,1\}$ is defined between a parent x_1 and a child x_2. $h^{x_1x_2} = 0$ if x_2 inherits the paternal haplotype of x_1, $h^{x_1x_2} = 1$ if x_2 inherits the maternal haplotype of x_1.*

Mendelian laws of inheritance impose constraints on *ps* and h variables for each parent-child pair at each locus. These constraints can be represented by a linear relationship of *ps* and h variables over the group $(Z_2, +)$ (where $0+0=0$, $0+1=1$, $1+1=0$). To process the constraints, [1] introduced the concept of locus graph. The original idea of [1] was to integrate edge constraints to construct a new subsystem that only consists of h variables first. Their algorithm will then solve the subsystem and use its solutions to solve the *ps* variables. We also record these constraints on locus graphs. However, instead of explicitly listing and solving the constraints on h variables, we use disjoint-set structures to collect, encode and thus examine the consistency of these constraints, which help us achieve a better time complexity result to obtain a general solution.

There are essentially two types of constraints on h variables in a locus graph L_i, *path constraints* and *cycle constraints*. For example, if p_i^s and p_i^t are pre-determined constants, we will have a path constraint on h variables, which is

$$\sum_{e_{xy} \in \widetilde{P_{v_s,v_t}}} h^{xy} = p_i^s + p_i^t + \sum_{e_{xy} \in \widetilde{P_{v_s,v_t}}} c_i^{xy} \tag{1}$$

where the right-hand side is a constant. A similar equation can be defined for a cycle constraint. By exploiting special features of the constraints on h variables, it is not necessary to explicitly list every

[1]Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland OH 44106, USA. Email: `xin.li2@case.edu, jingli@case.edu`

path and cycle constraint to check their consistency. We employ disjoint-set structures to detect and to check the consistency of constraints on h variables. For each locus graph L_i, we build a disjoint-set structure D_i to encode its connectivity information. We update the disjoint-set structure incrementally upon processing each edge constraint on a locus graph. Path constraints on a locus graph are detected during this process and will be stored in another disjoint-set structure D. The whole algorithm works on $m+1$ such disjoint-set structures, one D_i for each locus graph L_i and one D for encoding all path constraints. Cycles on a locus graph from a tree pedigree can only be generated within a nuclear family when it has multiple children. The algorithm first breaks all such short cycles through node splitting, which results in only path constraints for further processing. D_i will then be constructed from each locus graph L_i to recode the connectivity information and to detect path constraints. Processing of constraints and consistency check will be then performed, and a general solution of h variables will be encoded in the disjoint-set structure D. Solutions of ps variables will then be obtained. The algorithm can correctly output a general solution in near linear time $O(mn\alpha(n))$ on a tree pedigree without missing data. One of the advantages of the proposed algorithm is that it can be easily extended to the general cases of looped pedigrees and pedigrees with missing data (with some extra time).

3 Results

We have implemented the above algorithm and we have studied the performance of our program (denoted as ZRHC) under dierent settings (pedigree size, number of loci, missing rate, pattern of missing) and compare its performance with two representative programs Merlin [2] and PedPhase (the integer linear programming ILP algorithm in [3]). All three algorithms output exact the same set of haplotype congurations in our setting. Merlin and PedPhase.ILP scale exponentially with parameters. While ZRHC scales smoothly with all parameters, and the improvement over Merlin or PedPhase.ILP is from 10 to 10^5 folds for large pedigrees with large number of loci or high rate of missing (Figure 1).

References

[1] Xiao J., Liu L., Xia L., Jiang T. (2007) Fast elimination of redundant linear equations and reconstruction of recombination-free Mendelian Inheritance on a pedigree. In: *Proc. 18th Annual ACM-SIAM Symoposium on Discrete Algorithms* (SODA'07), 655–664.

[2] Abecasis G. R., Cherny S. S., Cookson W. O., Garden L. R. (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, 30(1):97–101.

[3] Li J. and Jiang T. (2004) An exact solution for finding minimum recombinant haplotype configurations on pedigrees with missing data by integer linear programming. In: *Proc. RECOMB'04*, 20–29.

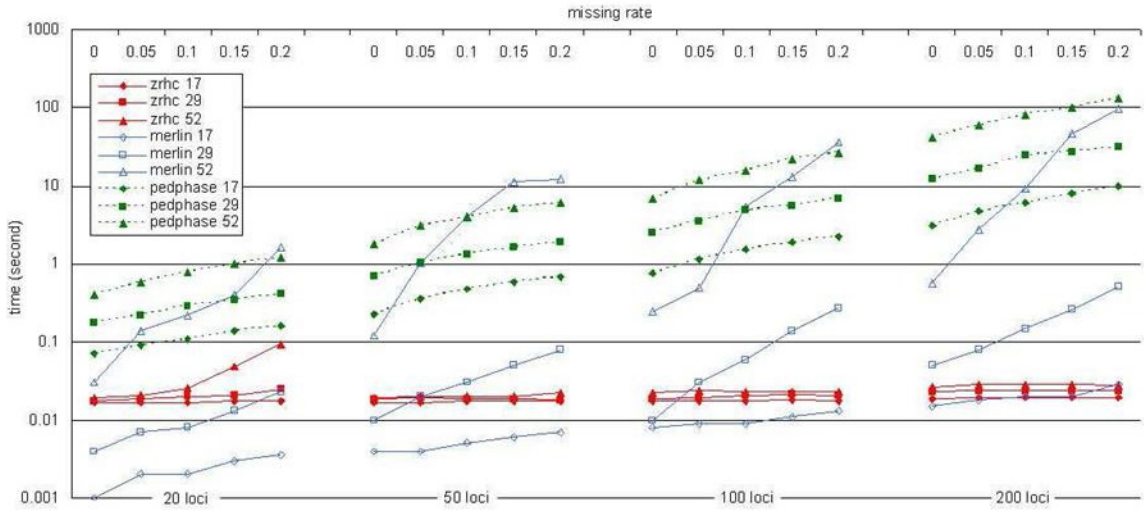


Figure 1: Comparison of running time (in seconds).

A Document Classification Strategy to Predict Protein Subcellular Localization Using Sequence Motifs and Evolutionary Information

Jia-Ming Chang, Emily Chia-Yu Su, Allan Lo, Hua-Sheng Chiu, Ting-Yi Sung, Wen-Lian Hsu[1]

1 Introduction

Protein subcellular localization is important for genome annotation, protein function prediction, and drug discovery. However, determination of subcellular localization using experimental approaches is time-consuming; thus, efficient prediction using computational approaches becomes highly desirable. We present a prediction method, PSLDoc (Protein Subcellular Localization prediction based on Document classification), which incorporates a probabilistic latent semantic analysis (PLSA) with a one-vs-rest support vector machine (SVM) model based on document classification techniques for both prokaryotes and eukaryotes.

Our method extracts biological features from gapped-dipeptides of various distance, where evolutionary information from the position specific score matrix is utilized to determine the weighting of each gapped-dipeptide. Then, the features are further reduced by PLSA and incorporated as input vectors for SVM classifiers. The accuracy of PSLDoc reaches 93.0% for Gram-negative bacteria proteins and 81.7% for human proteins in a five-fold cross-validation compared to previous results of 91.2% and 78.0%, respectively. Experiment results show that feature selection and reduction by document classification techniques can lead to a significant improvement in the prediction performance. Moreover, we demonstrate that PLSA automatically selects discriminating sequence motifs and greatly reduces the feature dimension without sacrificing the prediction accuracy.

Most notably, compared to similar approaches based on motif co-occurrences, PSLDoc achieves a much higher coverage because it starts with the examination of dipeptides and also considers the collocation of higher-order sequence motifs by PLSA feature transformation. Because of the generality of this method, it can be extended to more species or multiple localization sites in the future.

2 Figures and Tables

Table 1 shows the performance of PSLDoc, HYBRID [1], and PSORTb v2.0 [2] on PS1444 [3]. PSLDoc achieves the best performance of 93.01%, better than HYBIRD of 91.6% and PSORTb of 82.6%.

Loc. Sites	PSLDoc		HYBRID		PSORTb v2.0	
	Acc(%)	MCC	Acc(%)	MCC	Acc(%)	MCC
CP	94.96(94.24)	0.91(0.91)	95.00	0.89	70.10	0.77
IM	93.20(93.53)	0.94(0.94)	90.60	0.92	92.60	0.92
PP	89.13(89.13)	0.87(0.85)	88.80	0.84	69.20	0.78
OM	95.65(95.14)	0.95(0.94)	95.10	0.93	94.90	0.95
EC	90.00(87.37)	0.87(0.86)	85.30	0.87	78.90	0.86
Overall	93.01(92.45)	-	91.60	-	82.60	-

Table 1: Comparison of PSLDoc, HYBRID and PSORTb v.2.0 on the PS1444 data sets. The PSLDoc performance of incorporating a three-way data split procedure is indicated in the parentheses.

References

[1] Yu, C.S., Chen, Y.C., Lu, C.H., Hwang, J.K. 2006. Prediction of protein subcellular localization. *Proteins*, 64:643–651.

[2] Gardy, J.L., Spencer, C., Wang, K., et al. 2003. PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Research*, 31:3613–3617.

[3] Rey, S., Acab, M., Gardy, J.L., et al. 2005. PSORTdb: A protein subcellular localization database for bacteria. *Nucleic Acids Research*, 33:D164–D168.

[1]Bioinformatics Lab, Institute of Information Science, Academia Sinica, Taipei, Taiwan. Email: `hsu@iis.sinica.edu.tw`

Statistical Analysis of Macrophage Cell Morphology after Microtubule Disruption

A. Ng,[1,2] J.C. Rajapakse,[1,2] J.G. Evans,[3] R. Welsch[4]

1 Introduction

High Content Screening (HCS) is a high throughput technology that applies sophisticated image processing algorithms to analyze cell images generated by automatic fluorescence microscopy [5]. HCS has gained popularity as a systemic approach in the large scale study of biological systems. The requirement for analyzing and mining data generated from HCS extends well beyond the present generation of informatics tools [3]. The real benefit of utilizing HCS in systems biology research can be enhanced via the use of statistical and machine learning tools.

We study the dose dependent response of IC-21 macrophages to a microtubule disrupting drug, demecolcine [2]. Cell morphology readings are taken over 65 parameters that cover intensity, texture and shape over three fluorescence dyes that demarcate the cytoplasm (CMFDA), nucleus (Hoechst) and F-actin (Texas red phalloidin). Cells are treated over 48 drug concentrations (16 attoM to 1 microM) with dimethyl sulfoxide (DMSO) used as a control.

2 Methodology

We use the Kolomogorov-Smirov (K-S) statistics to quantify differences between demecolcine and DMSO treated cell populations. The K-S statistic for each morphological feature per drug concentration yields a 65 x 48 heat map which represents the drug response profile of demecolcine.

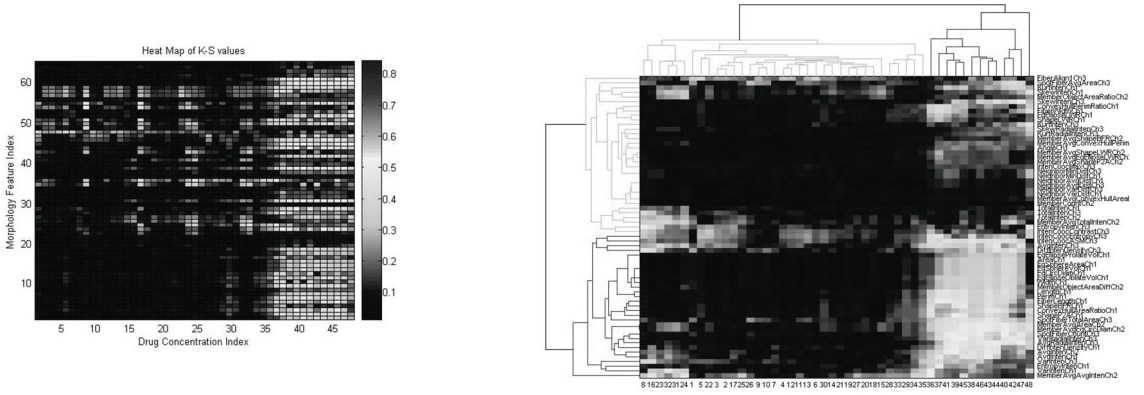


Figure 1: Heat Map of K-S values before and after bi-clustering. The 65 x 48 values obtained from each morphological feature per drug dosage. The map is color coded to show the morphological parameters that did not change (blue), changed moderately (orange) and changed extensively (red) as a result of demecolcine treatment. After bi-clustering (right map), the significant morphological features and the drug concentrations which they occurred clustered as the block on the lower right corner of the heat map.

Distinctive patterns in the K-S Heat Map can be identified by simultaneously clustering features and drug dosages with bi-clustering. 31 features at 13 drug dosages are identified as significantly different from the control. The Primary Effective Concentration (PEC) [4] of demecolcine occurs at Concentration 36 (0.026M). This concentration reflects the point at which many of the morphology descriptors become different from the control values.

[1]Singapore MIT Alliance, Nanyang Technological University, Singapore 637460. Email: R060002@ntu.edu.sg
[2]BioInformatics Research Centre, Nanyang Technological University, Singapore 637553. Email: ASJagath@ntu.edu.sg
[3]Whitehead BioImaging Center, MIT, MA 02139, USA. Email: jgevans@wi.mit.edu
[4]Sloan School of Management, MIT, MA 02142, USA. Email: rwelsch@mit.edu

Z-scores quantify the significant differences between the general population of cells with post-PEC cells. A high absolute Z-score ranks an important feature. A selection of the 31 significant features are tabulated.

No	Feature Name	μ_{36-48}	σ_{36-48}	μ_{1-48}	σ_{1-48}	Z score
27	EntropyIntenCh1	8.53	1.02	9.73	0.91	-185.24
24	VarIntenCh1	1366.39	312.88	931.90	367.76	165.06
23	AvgIntentCh1	1498.49	324.40	1139.11	371.24	135.25
6	LengthCh1	39.45	11.96	51.07	19.28	-84.18
2	PerimCh1	122.53	38.81	165.68	73.34	-82.21
7	WidthCh1	27.56	8.18	34.67	12.65	-78.73
35	MemberAvgAvgIntenCh2	1303.42	274.46	941.83	256.59	196.89
43	AvgIntenCh2	334.04	112.70	245.92	101.59	121.17
38	MemberAvgEqCircDiamCh2	13.99	3.07	16.47	3.49	-99.06
30	MemberAvgAreaCh2	101.20	82.12	222.71	105.08	-81.79
50	VarIntenCh3	531.65	213.42	301.79	174.14	184.41
60	DiffIntenDensityCh3	84.13	28.60	62.25	25.87	118.19
49	AvgIntenCh3	468.58	152.00	372.04	127.23	106.01
44	SpotFiberCountCh3	64.73	45.81	114.04	77.55	-88.84

Table 1: List of a some significant features describing the change in morphology after demecolcine treatement. In particular, cytoplasm and nuclear sizes decrease, while fluorescence increase. F-actin spots are fewer and its fluorescence brighter.

We incorporate our morphological findings with current literature [1] to develop a model of the macrophage cytoskeleton upon demecolcine treatment (see Figure 2).

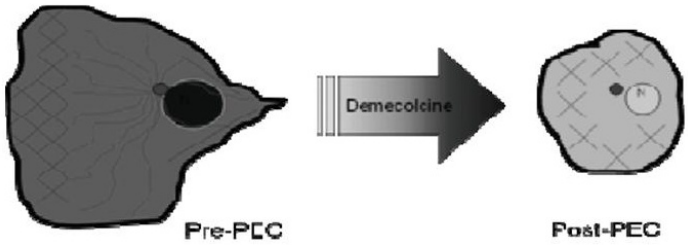


Figure 2: Model of the main cytoskeletal descriptors upon Demecolcine treatment. The Pre-PEC shows microtubles (pink) extending towards the protruding edge with actin fibers (red), nucleus (blue). The different shades of green and blue are meant to depict fluorescence changes.

3 Conclusion

We have demonstrated use of statistical tools in a simple but novel way of identifying and selecting important features in a HCS drug response profile. With these selected features, we propose a model describing macrophage cytoskeletal behavior under demecolcine treatment.

References

[1] Etienne-Manneville, S. 2004. Actin and microtubules in cell motility: which one is in control? *Traffic*, 5:470–477.

[2] Evans, J.G., and P. Matsudaira. 2007. Linking microscopy and high content screening in large-scale biomedical research. *Methods Mol Biol*, 356:33–38.

[3] Giuliano, K.A., P.A. Johnston, A. Gough, and D.L. Taylor. 2006. Systems cell biology based on high-content screening. *Methods Enzymol*, 414:601–619.

[4] Perlman, Z.E., M.D. Slack, Y. Feng, T.J. Mitchison, L.F. Wu, and S.J. Altschuler. 2004. Multidimensional drug profiling by automatedmicroscopy. *Science*, 306:1194–1198.

[5] Taylor, D.L. 2007. Past, present, and future of high content screening and the field of cellomics. *Methods Mol Biol*, 356:3-18.

Fragment-Based Analysis of Protein-Ligand Interactions Using Localized Stereochemical Features

Reetal Pai,[1] James Sacchettini,[2] Thomas Ioerger[3]

1 Introduction

The binding affinity between a protein and its cognate ligand is determined by their steric and chemical complementarity. Yet, modelling binding site patterns and using them to predict cognate ligands remains challenging. Computational analyses of protein structure-function relationships have traditionally been based on sequence homology, fold family analysis and 3D motifs/templates [2]. Despite the successes of these approaches, they are unable to capture similarities between active sites that span multiple fold families despite catalyzing the same reaction (convergent evolution). In a recent paper [1], the authors observed that significant variations exist in the shape of active sites binding the same ligand due to the flexibility observed in larger ligands (due to the number of internal degrees of freedom). This flexibility makes it additionally difficult to capture the patterns of protein-ligand interaction.

In this work, we use a novel fragment-based approach to limit the effect of ligand flexibility on active site analysis. We also extend previous feature-based analyses of active sites by defining a system of localized geometric and electrostatic descriptors that identify localized patterns of protein-ligand interactions. Singular Value Decomposition is used to identify linear combinations of features with maximum information content which are then used to compute the class conditional probability density distribution of active sites using kernel density estimation. In the case of multi-fragmented ligands, this methodology is extended by combining predictions based on a graphical model. We successfully tested our algorithm on a database that contained examples of adenine, citrate, nicotinamide, phosphate, pyridoxal and ribose binding proteins with over 75% accuracy. We also tested our fragment-based approach on an AMP binding site and accurately identified the position of the ligand within the active site.

2 Application to Multi-Fragment Ligands

Our database is composed of fragments of larger ligands and therefore, to predict the binding of larger ligands we analyze the entire active site of an unliganded protein in a piecewise manner and then combine the piecewise analysis to recognize the native ligand that fits the entire site. The analysis begins with breaking the active site into overlapping regions by considering each active site vertex as the center of a region and all other surface vertices within a chosen uniform radius are considered to be part of this region. Localized sterochemical features are then computed for each of the regions and these vectors are then transformed onto a reduced-dimension SVD space based on the feature vectors from our database. These reduced-dimension vectors are used to compute the posterior probability of each ligand class given the feature vector at active site vertex j, i.e. $P(C_i|x_j)$, using kernel density estimation.

In addition to the fragment database, a database containing multiple examples of all of the corresponding larger ligands (AMP, ADP, ATP, NAD, PLP, etc.) is also created. These larger ligand examples are used to create a statistical model that describes the relationship between the positions of the centers of the corresponding fragments using simple distance constraints. In order to complete the analysis of the larger ligand, we model complex ligands using Bayesian graphical models where the nodes N are the ligand fragments and the edges E denote the connectivity between fragments. Absent edges are assumed to represent conditional independence. The probability of any multi-fragment ligand can be written as a joint PDF:

$$P(f_i, \ldots, f_n) = \prod_{f_i \in N} P(f_i) \prod_{(f_j, f_k) \in E} P(f_j|f_k) \quad (1)$$

An example graphical model for the ligand AMP is shown in Figure 1.

[1]Department of Computer Science, Texas A&M University. Email: reetalp@cs.tamu.edu
[2]Department of Biochemistry & Biophysics, Texas A&M University. Email: jim.sacchettini@gmail.com
[3]Department of Computer Science, Texas A&M University. Email: ioerger@cs.tamu.edu

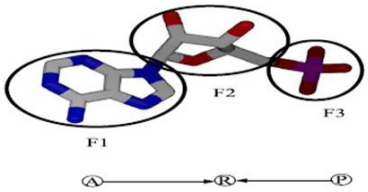


Figure 1: The three fragments of adenine, ribose and phosphate that comprise the ligand AMP and a graphical model showing the positional dependence between the centers of these fragments.

3 Results and Discussion

Table 1 shows the various ligand fragments used in this study. The examples within each ligand class were chosen so as to create a diverse dataset containing examples belonging to various fold families with very low sequence homology to each other. The third column in Table 1 shows the diversity in fold families within the dataset and the fourth column shows the average homology of the examples of each fold to the other members of the ligand family (less than 30% in all cases).

The active site surface can be defined as all those protein-molecular surface atoms that are in contact with the ligand. This definition requires knowledge of ligand coordinates unavailable in the case of hypothetical proteins. To ensure that our methodology was extendable to the study of hypothetical proteins, we tested the accuracy of the feature-based approach using active site surfaces created based on known ligand coordinates as well as using uniform radius active sites centered on atom closest to ligand center. The fifth and the sixth columns of Table 1 show the efficacy of our novel localized stereochemical acive site descriptors designed to capture spatial information about the pocket shape and electrostatic nature on the active sites generated using the two different techniques.

The seventh and the eighth columns in Table 1 show that in the case of all ligand classes there is a 30% or greater drop in classiffication accuracy when only geometric or only electrostatic features are used except in the case of phosphate binding examples. When only geometric features are used, there is a clear distinction between active sites that bind phosphate and those that do not, leading to a 100% accuracy in the classification of phosphate binding sites. Despite this anomaly, these results clearly show that neither shape nor electrostatic descriptors alone are sufficient to describe the active site patterns and that it is necessary to combine these features to capture the binding patterns observed among the diverse active sites binding the same ligand.

We applied the fragment-based analysis to 2AK3 (adenylate kinase that binds AMP) and 2LYZ (lysozyme that does not bind AMP). Despite the inaccuracies in shape descriptions with the use of fragment-based analysis as well as uniform-radius active site definitions, we were able to define a ligand position within the active site of 2AK3 very close to the original position. We were also able to correctly distinguish it from 2LYZ, a non-AMP binding protein. This is highly encouraging for future applitcations to functional annotation of hypothetical proteins. Additionally, this piecewise analysis of a large active site has higher probability in helping in the identification of other inhibitors and the design of other small-molecules to interact with the active site.

Ligand Name	# Fold Families	# Examples Within Class	Ave Homology Betw Families	Accuracy Contact Surface	Accuracy Uniform Radius Surface	Accuracy Only Geometric Features	Accuracy Only Shape Features
Adenine	19	55	7.9%	78%	69%	45%	52%
Citrate	12	16	8.7%	81%	56%	44%	38%
Nicotinamide	11	55	7.3%	84%	64%	56%	40%
Phosphate	23	55	6.9%	91%	96%	100%	45%
Pyridoxal	6	55	10.2%	87%	67%	36%	47%
Ribose	22	55	8.5%	78%	64%	44%	42%

Table 1: Description of fragment database and analysis of classiffication accuracy.

References

[1] Kahraman A., Morris RJ., Laskowski RA., Thornton JM. 2007. Shape variation in protein binding pockets and their ligands. *J Mol Biol*, 368(1):283–301.

[2] Laskowski RA., Watson JD. Thornton JM. 2005. Protein function prediction using local 3D templates. *J Mol Biol*, 351:614–626.

QSAR Studies of Anti Influenza Neuraminidase Inhibitors [Oseltamivir]

Avinash Kumar S., Namit Bharija[1]

1 Introduction

Influenza, one of the most common infectious diseases of mammals in general and humans in particular, is caused by an RNA virus of the paramyxoviridae family (The influenza virus). 'Flu', as it is more commonly known, spreads in seasonal epidemics that may sometimes develop into deadly pandemics, wreaking havoc on a global scale. Each flu pandemic so far has been caused by the appearance of a new strain of the virus. The appearance of the new bird flu strain (H5N1) has reawakened fears of an impending flu pandemic.

Due to the high mutation rate of the virus, efforts to produce a vaccine against the disease have not been successful. Combating the flu is currently dependent on anti-influenza drugs that are being commercially produced. Neuraminidase inhibitors are a one of the most successful categories of anti-influenza drugs that have been developed. Neuraminidase is an enzyme that helps the virus to enter into host cells during the infection. Use of neuraminidase inhibitors has been shown to greatly reduce the infection rate due to influenza virus.

QSAR [Quantitative Structure Activity Relationship] analysis is a way to quantitatively correlate the biological activity of a molecule to its structure. (i.e. find a relationship between the structure and function). To better understand the mechanism of action of neuraminidase inhibitors and to help in designing better and efficient drugs against influenza, a QSAR analysis of the interaction between Neuraminidases N1 & N2 (two viral enzymes that have been implicated in most influenza epidemics) and the commercial drug Oseltamivir [TamifluTM] was performed.

3D QSAR is a related method that gives quantitative relationship based not on the values of various molecular descriptors but on the 3D interaction between the ligand and receptor. This too was performed for the chosen enzymes and drug analogues. The COMFA method was used to perform 3D QSAR.

2 Software and Files

The structural analogues were chosen so as to give a total picture of the factors impacting the interaction, based on docking studies of Oseltamivir with the two enzymes. Structural analogues of Oseltamivir were obtained from the Pubchem database or created using Chemsketch. Crystal structures of the neuraminidase enzymes were sourced from PDB database. Various biological activity values that were analysed during the course of the study (IC50, ClogP, XlogP) were all calculated using Quantum ver 3.0 [1].

Various descriptors like topological and 2D autocorrelation descriptors for the training set were calculated by e-dragon server. A multiple regression analysis was performed and various QSAR models were obtained. All the analogues were then used for performing a 3D-QSAR [2] analysis by CoMFA method. The fit atom based alignment yielded best predictive CoMFA model .The contour maps obtained from 3D-QSAR studies were appraised for the activity trends of the molecules analyzed

[1]School of Biotechnology, Chemical and Biomedical engineering, VIT University, Vellore, India.

3 Figures and Tables

Mol ID	IC50	pIC50	Wap	D/Dr06
480256	0.0475	1.323306	1056	55.486
480257	0.0177	1.752027	1248	59.914
480260	0.0214	1.669586	1694	69.137
480262	0.00721	2.142065	2426	83.076
490476	0.00929	2.031984	2711	86.613
490477	0.01428	1.845272	1689	69.698
493853	0.0194	1.712198	2326	79.123
493857	0.0732	1.135489	3948	99.71
493859	0.00339	2.4698	2158	78.321
493872	0.0161	1.793174	2175	77.573
493873	0.03557	1.448916	2492	82.329
493877	0.01447	1.839531	2113	77.432
505918	0.01078	1.967381	1470	64.525
505919	0.01633	1.787014	2235	78.792
505922	0.02069	1.68424	2157	78.359
505929	0.01563	1.806041	2113	77.432
5464654	10.2751	-1.01179	5236	183.743
6481599	0.3549	0.449894	1470	64.525
6481600	0.0273	1.563837	2326	79.123
6483690	0.0958	1.018634	3065	89.296
9926260	0.0143	1.844664	2175	77.573
9967681	0.01601	1.795609	3044	91.224
10357442	0.00853	2.069051	2158	78.321
10640758	0.00664	2.177832	3188	91.812
10713139	0.0885	1.053057	3410	95.835
10763205	15.956	-1.20292	4368	108.883
10881879	0.016	1.79588	2714	86.613
11441321	2.99	-0.47567	4988	114.53
11630102	0.02136	1.670399	1246	60.335
15956756	0.01383	1.859178	1667	68.954

Significant Descriptor Values

1. WAP regression model: pIC50=2.925234 - 0.000596(wap)
2. D/Dr06 regression model: pIC50=1.433855- 0.00182(D/Dr06)

References

[1] Min-Jie Li, Chen Jiang, Ming-Zong Li, Tian-Pa You. QSAR studies of 20(S)-camptothecin analogues as antitumor agents. *Journal of Molecular Structure: THEOCHEM*, 723:165–170, 2005.

[2] Tabish Equbal, Om Silakari, Muttineni Ravikumar. Three-dimensional quantitative structure activity relationship (3D-QSAR) studies of various ether analogues of Farnesyltransferase Inhibitors. *Internet Electronic Journal of Molecular Design*.

De Novo Design of Peptides: Potential Vaccines against the Influenza A Virus

Sandro Andreotti,[1] Jürgen Kleffe,[2] Paul Wrede[3]

1 Introduction

The human influenza A virus causes about 100 000 deaths per year worldwide [6]. Until today the design of an effective vaccine failed because of the high variability of the virus. Especially the two viral surface proteins haemagglutinin (HA) and neuraminidase (NA) mutate very rapidly. These two proteins are recognized by immunoglobulins of the adaptive immune system. Current influenza A vaccines depend on the production of inactive sub-virions in embryonated eggs. For each season a special vaccine must be produced in advance. For 30 million people 90 million eggs are required to get enough material. The production takes about 6 to 8 month. A concept is suggested for the rational *de novo* peptide design to gain subtype independent vaccines which can be used each year. An advantage of using peptides in vaccination is their cheap production in kilogram scale and high stability for years.

2 Concept

A novel idea of vaccination against influenza A virus is the stimulation of the T-cell mediated immune response. When a virus infects a cell the viral proteins are fragmented and via the intra-cellular transport system nonameric peptides are bound to MHC I molecules which are displayed on the cell surface. This peptide-MHC I complex can be recognized by a CD8 T-cell. In case such peptide is unknown to the T-cell it destroys the infected cell immediately. Peptides recognized by T-cells are called T-cell epitopes.

Our concept includes the computer-based rational design of T-cell epitopes which are derived from highly conserved regions of the viral proteins [3]. To decide which nonamer peptide can be used as influenza A vaccine a thorough search for all known nonamers of the viral proteome is done with the fast program ClustDBP [4]. In addition a comparison with all unique nonamers of the human proteome revealed a disjoint distribution set of viral and human nonamers. Both sets have only 4 nonamers in common, three occur in the viral protein NA and one in PB2. The total number of different nonamers for human is: 9 109 196; for influenza A: 78 288. For avian influenza A viruses 153 545 unique nonamers exist, not surprising human influenza A and avian influenza A virus have 31 190 nonamers (HA: 6858; NA: 5954) in common. Both sets of human and viral nonamers represent only a very small fraction of the total nonamer sequence space with 5.120 000 000 000 peptides. For the human influenza A proteins we analyzed also the number of octamers and hexamers. Total number of octamers: 11 716 493; hexamers: 11 770 815, unique octamers: 69 708; unique hexamers: 52 748. There are 102 octamers and not surprising 12 387 hexamers in common with the total human peptidome.

In order to design influenza A vaccines a number of nonamer T-cell epitopes should be derived from the conserved regions of several viral proteins. We analyzed for all nonamers the frequency of occurrence in each of the ten viral proteins (Table 1). Originally this table has 78 288 rows of which we show only 12 for space reason. The peptide FGAIAGFIE occurs in 2819 from 2830 HA proteins. It does not exist in any other of the viral proteins. In another case the peptide PFLDRLRRD occurs in 2650 NS 1 proteins and in only one NS2 protein while the peptide EQITFMQAL is found in 2619 NS2 proteins and in one NS1 protein. Two proteins occur as alternative splice variants MP1 and MP2 as well as NS1 and NS2. In both cases a single peptide was found almost in all MP 1 and MP 2 proteins while 65% of the NS1 and 66% of the NS2 proteins contain the same peptide.

A complete analysis of these data will be presented on the poster. Our results show that a sufficient number of conserved peptides not occurring in the human genome exist. Currently we investigate similarity searches on the basis of physicochemical properties of the peptides. Selected peptides are used as training data for improving our current T-cell epitope prediction algorithms [1, 2]. The rational *de novo*

[1]Charité-Universitätsmedizin Berlin, Institut für Molekularbiologie und Bioinformatik, Arnimallee 22, D-14195 Berlin.
[2]Email: juergen.kleffe@charite.de
[3]Email: paul.wrede@charite.de

design will be a combination of evolutionary algorithms in combination with experimental immunological testing of the designed peptides [5, 7, 8, 9]. Each set of peptides is analyzed by using the fast matching algorithm ClustDBP for similarity with the human peptidome. This will be a first step in selecting out most likely candidates for cross reaction.

References

[1] Bredenbeck, A., Losch, F., Sharav, T., et al. 2005. Identification of non-canonical melanoma-associated T-cell epitopes for cancer immunetherapy. *Journal of Immunology*, 174, 6716–6724.

[2] Filter, M., Eichler-Mertens, M., Bredenbeck, A., et al. 2006. A strategy for the identification of canonical and non-canonical MHC I-binding epitopes using an artificial neural network based epitope prediction algorithm. *QSAR & Combinatorial Science*, 25:350–358.

[3] Ghedin, E. et al. 2005. Large scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, 437:1162–1166.

[4] Kleffe, J., Möller, F., Wittig, B. 2006. ClustDB: A high-performance tool for large scale sequence matching. In: *DEXA Workshops*, pages 196–200.

[5] Schneider, G., Wrede, P. 1998. Artificial neural networks for computer-based molecular design. *Progress in Biophysics and Molecular Biology*, 70:175–222.

[6] World health Organization. Cumulative number of confirmed human cases of avian influenza A/(H5N1) reported to WHO.

[7] Wrede, P., Filter, M. 2007. Bioinformatics: Algorithms for T-cell epitope based vaccine design. In: MFG Schmidt, *Influenza Viruses, Facts and Perspectives*, Berlin.

[8] Wrede, P., Schneider, G. 1994. *Concepts in protein engineering and design*. Berlin: Walter de Gruyter.

[9] Wrede,P., Landt, O., Klages, S., et al. 1998. Peptide design aided by neural networks: Biological activity of artificial signal peptidase I cleavage sites. *Biochemistry*, 37:3588–3593.

Viral protein:	HA	MP1	MP2	NA	NP	NS1	NS2	PA	PB1	PB2
Length	565	249	95	467	497	228	120	715	757	758
#. sequences	2830	2839	2827	3409	2635	2700	2638	2508	2488	2522
FGAIAGFIE	2819	0	0	0	0	0	0	0	0	0
FVQNALNGN	0	2825	1	0	0	0	0	0	0	0
PESMREEYR	0	0	2723	0	0	0	0	0	0	0
ILRTQESEC	0	0	0	3377	0	0	0	0	0	0
MIWHSNLND	0	0	0	0	2634	0	0	0	0	0
PFLDRLRRD	0	0	0	0	0	2650	1	0	0	0
EQITFMQAL	0	0	0	0	0	1	2619	0	0	0
HLRNDTDVV	0	0	0	0	0	0	0	2508	0	0
AIATPGMQI	0	0	0	0	0	0	0	0	2488	0
KAVRGDLNF	0	0	0	0	0	0	0	0	0	2521
MSLLTEVET	0	2709	2703	0	0	0	0	0	0	0
DSNTVSSFQ	0	0	0	0	0	1778	1754	0	0	0

Table 1: Frequency of occurrence of nonamers in each of the ten human influenza A viral proteins.

Size-Specific and Brightness-Weighted Cell Tracking in 2D images

Merlin Veronika,[1,2] James G. Evans,[3] Paul Matsudaira,[1,3,4]
Roy E. Welsch,[1,5] Jagath C. Rajapakse[1,2]

1 Introduction

One of the major challenges of biomedical research in the post genomic era is the unraveling of the spatiotemporal relationships of complex biomolecular systems [5]. Naturally this involves acquisition of timelapse image series and tracking of objects over time. From image analysis point of view, a distinction can be made between tracking of single molecules (or complexes) and tracking of entire cells. A number of tools are available for studying the dynamics of proteins based on fluorescent labeling and time-lapse imaging, such as fluorescence recovery after (and loss in) photo bleaching (FRAP and FLIP respectively), but these methods yield only ensemble average measurements of properties. More detailed studies into the different modes of motion of subpopulations require single particle tracking [3, 4] which aims at motion analysis of individual proteins or microspheres. We propose a brightness weighted centroid method for size-specific tracking of cells in time-series of two dimensional images. The method consists of two parts: segmentation of cells using a level-set method and tracking the centroid of the cells by Euclidean distance measure. The efficiency of this method can be improved by increasing the area of the cells under consideration.

2 Methodology

The segmentation method [1] can be described as a minimization of an energy based-segmentation. It implements Mumford-Shah functional via a level set function for bimodal segmentation. The segmentation is performed by an active contour model which uses the information inside regions rather than the intensity gradients along the edges. Segmented regions are represented via a level set function Φ which minimizes an energy funtional. The optimal curve would separate the interior and exterior with respect to their relative expected values. The local maxima is identified to pixel level accuracy by gray scale dilation and further approximated to the geometric center by measuring the offset using a pre-determined window (Fig. 1). Having located the objects in the sequence, the locations are matched in each image with the corresponding locations in the later image to produce the trajectories [2] given by

$$\rho(r,t) = \sum_{i=1}^{N} \delta(r - r_i(t))$$

where $r_i(t)$ is the location of the ith object in the field of N objects at time t.

3 Experimental Results

We demonstrate our method on series of six time-lapsed microscopic images taken from mouse microphage IC21 cell lines (Fig. 2). The data consists of approximately 475 cells per frame which differ in size, shape and volume. This method is suitable for tracking cells of specified size, i.e. of given diameter. From the above method we conclude that the dataset consists of approximately 330 cells which are less than or

[1]Computation and Systems Biology Programme, Singapore-MIT Alliance. Email: merlin@pmail.ntu.edu.sg
[2]BioInformatics Research Centre, Nanyang Technological University, Singapore. Email: asjagath@ntu.edu.sg
[3]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts. Email: jgevans@wi.mit.edu
[4]Department of Biology and Division of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts. Email: matsudaira@wi.mit.edu
[5]Sloan School of Management, Massachusetts Institute of Technology, Cambridge Massachusetts. Email: rwelsch@mit.edu

equal 12μ and travel not more than 25μ. Analyzing the velocity distribution, 70% of the cells travel with a velocity of $18\mu/h$.

References

[1] Chan, T.F and Vese, L.A., 2001. Active Contour without Edges. *IEEE Transactions on Image Processing*, 10(2):266–277.

[2] Crocker, J.C and Grier, D.G. 1996. Methods in digital video microscopy for colloidal studies. *Journal of colloidal and Interface Science*, 179:298–310.

[3] Qian, T and Sheets, M.P. 1991. Single particle tracking: Analysis of diffusion and flow in two-dimensional systems. *Biophysical Journal*, 60(4):910–921.

[4] Saxton, M.J and Jacobson, K. 1997. Single paticle tracking: Applications to membrane dynamics. *Annual Review of Biophysics and Biomolecular Structure*, 26:373–399.

[5] Tsien, R.Y. 2003. Imagining imaging's future. *Nature Reviews Molecular Biology*, 4(1):S16–S21.

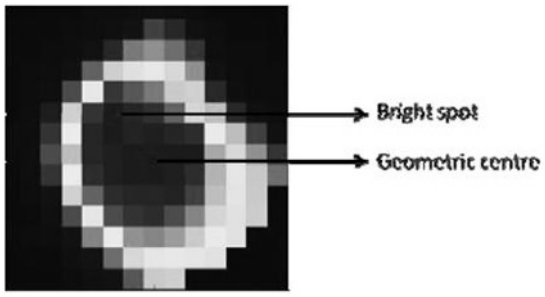


Figure 1: The location of object's centroid calculated from the local maxima.

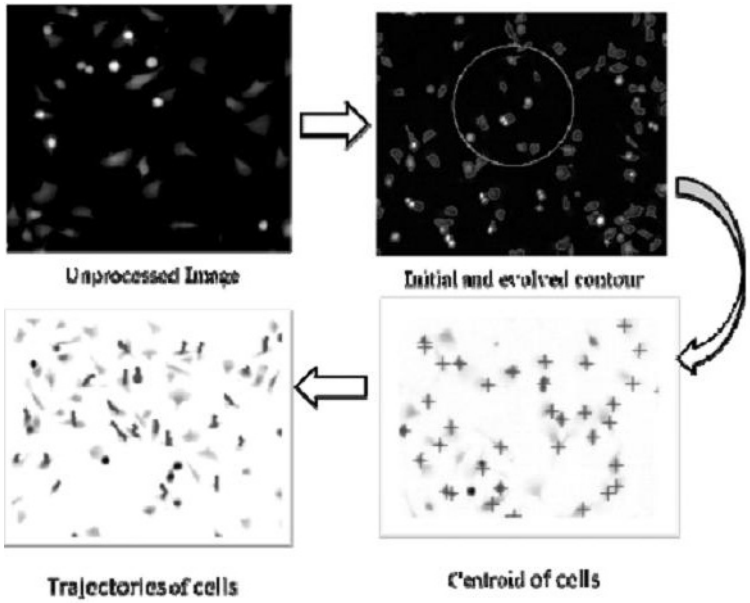


Figure 2: Results of segmentation and tracking cells using level set and brightness weighted approach.

EVOG: Evolution Visualizer for Overlapping Genes

Chi-Yong Cho,[1] Dae-Soo Kim,[2] Jae-Won Huh,[3] Heui-Soo Kim,[4]
DoHoon Lee,[5] Hwan-Gue Cho[6]

1 Introduction

Increasing the number of sequenced genomes has come a corresponding propagation of computational tools for analyzing, viewing, comparing, searching, and annotating genome sequences [1, 2, 3, 4, 5]. The UCSC Genome Browser Database(GBD) provides integrated sequence and annotation data for a large collection of vertebrate and model organism genomes [7]. The UCSC data are very useful but they have too many additional information to handle a few focussed information such as only revealing overlap genes among genomes.

Recently, increasing numbers of sense-antisense transcripts and overlapping genes have been identified in a variety of eukaryotic organisms using large-scale genome analysis. Many overlapping genes and sense-antisense transcripts have been indented in the genomes of eukaryotic, some of which have been reported to have functional roles, but their evolutionary origin is not clearly understood. We systematically analyzed all overlapping genes and sense-antisense transcripts in the eukaryotic genomes. In particular, careful comparisons were made for the othologous genes that are overlapped and sense-antisense transcripts in the various species.

So we developed a evolutionary visualization tool EVOG(Evolution Visualizer for Overlapped Genes) to visualize and analyze overlapped regions from the UCSC annotation data. This tool is simple and easy to control parameters for finding overlapped genes and sense-antisense transcripts in whole genomes.

2 Evolution Visualizer for Overlapping Genes

The UCSC Genome Browser [7] is a very nice and famous tool for analyzing and visualizing various information related genomes. Too much information gives us obstacles to understand what we should notice features from the biological data. We are interested in finding and analyzing the implication of sense-antisense transcripts and overlapping genes expression for human evolution and disease. The overlapped region may induce malfunction of genes when they have to express. To support this work, we developed searching and visualizing tool EVOG for finding overlapped genes and sense-antisense transcripts.

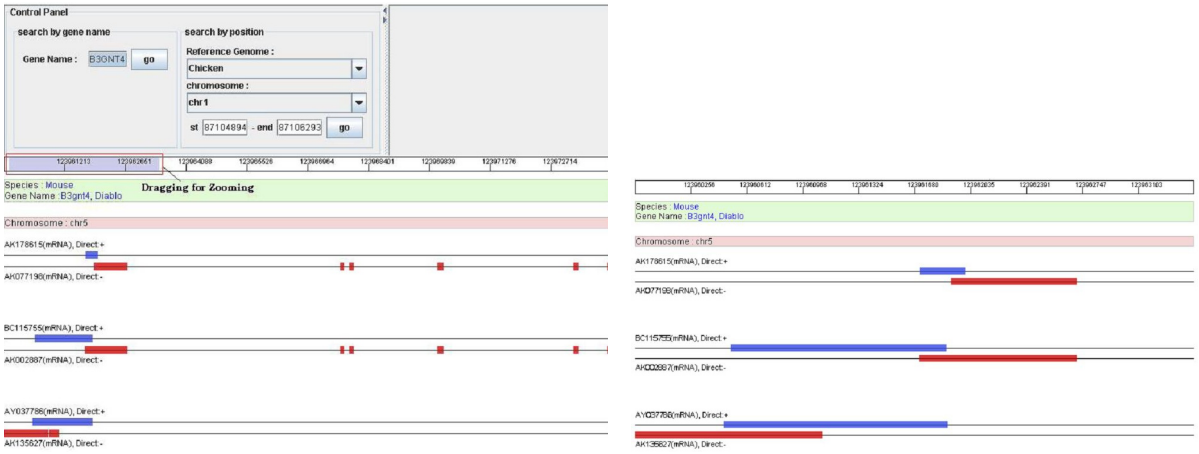
The EVOG use the UCSC annotation data as input and reports the overlapped genes among the whole genomes. Figure 1(a) shows a result of EVOG. We can get a whole overlapped genes among given 10 genomes by just gene name. For more detail information of special region shown in red box, EVOG supports dragging a region to zoom in and its zooming in-out in Figure 1(b). Gene B3GNT4 is overlapped with gene Diablo in Chromosome chr5 of Mouse.

The EVOG has the following features :

- Displaying all overlapped genes in whole genomes.
- Simple interface.
- Selecting genome, chromosome, and search area.

[1]School of Computer Science and Engineering, Pusan National University, Busan 609-735, Korea. Email: jane@blackhole.org
[2]PBBRC, Interdisciplinary Research Program of Bioinformatics, College of Natural Sciences, Pusan National University, Busan 609-735, Korea. Email: kds2465@pusan.ac.kr
[3]Division of Biological Sciences, College of Natural Sciences, Pusan National University, Busan 609-735, Korea. Email: primate@pusan.ac.kr
[4]Division of Biological Sciences, College of Natural Sciences, Pusan National University, Busan 609-735, Korea. Email: khs307@pusan.ac.kr
[5]School of Computer Science and Engineering, Pusan National University, Busan 609-735, Korea. Email: dohoon@pnu.edu
[6]School of Computer Science and Engineering, Pusan National University, Busan 609-735, Korea. Email: hgcho@pnu.edu

- Displaying additional annotation information of around overlapped regions.
- Adaptive zoom-in for rapid scale via dragging region.



(a) (b)

Figure 1: Interface of EVOG. (a) Up left of screen is control panel for setting gene name and parameters that can be selected by genome, chromosome, and position. The lower part shows the overlapped position, gene and genome name, and chromosome. B3GNT4 is overlapped with gene Diablo in Chromosome chr5 of Mouse. Red box is dragged area for zooming. (b) Zoom-in result of (a).

3 Discussion

We developed a intuitive and simple visualization tool for analyzing and displaying overlapped genes and senseantisense transcripts in whole genomes. The system also perform adaptive zooming, and its function is very useful to handle huge data such as genome. We use 10 vertebrate genomes, Human, Mouse, Rat, Chicken, Cow, Drosophila, Gambiae, Zebrafish, Xenopus tropicalls, and Chimp. From the those overlapped genes, we will systematically analyze evolutionary relationship among species in the future. Furthermore EVOG will help us gain insight into the implications of sense-antisense transcripts and overlapping genes expression for human evolution and disease. EVOG is available on `http://164.125.34.87/~evog`.

References

[1] Schwartz, S., Zhang, Z., Frazer, K. A., et al. 2000. PipMaker: A web server for aligning two genomic DNA sequences. *Genome Research*, 10:577–586.

[2] Ovcharenko, I., Nobrega, M. A., Loots, G. G., Stubbs, L. 2004. ECR Browser: A tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Research*, 32, web server issue.

[3] Brudno, A. P., Salamov, A., Cooper, G. M., et al. 2004. Automated whole-genome multiple alignment of rat, mouse, and human. *Genome Research*, 14:685–692.

[4] Bailey, J. A., Eichler, E. E. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Reviews Genetics*, 17:552–564.

[5] Gans, J. D. and Wolinsky, M. 2007. Genomorama: Genome visualization and analysis. *BMC Bioinformatics*, 8:204.

[6] Dahary, D., Elroy-Stein, O., Sorek, R. 2005. Naturally occuring antisense: Transcriptional leakage or real overlap? *Genome Research*, 15:364–368.

[7] Karolchik, D. et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Research*, 36:D773–D779.

A Study of Microsatellites Dominating Mammalian Size Variation

Meng-Chang Hsiao,[1] Chien-Ming Chen,[1] Tun-Wen Pai,[1,*]
Wen-Shyong Tzou,[2] Ron-Shan Chen[3]

1 Abstract

Microsatellites, also referred to as variable number of tandem repeats or simple sequence repeats (SSRs), are polymorphic loci in the genome that consist of repeating units of 1–6 base pairs in length and play a crucial role in genome mapping and various genetic studies. The Insulin-like growth factor 1 (IGF1) gene is a highly conserved polypeptide which regulates growth and metabolic functions in several vertebrate species. In a recent study, alleles of a dinucleotide $(CA)_n$ microsatellite appeared within different frequencies and located at the promoter regions of the IGF1 gene for variant sizes of dogs are unveiled, and the allelic forms of the microsatellites are significantly associated with adult body size. Base on these discovered results, we have developed a system employing series genomewide comparative genomics analyses and tried to efficiently identify whether the important microsatellites are also located in the promoter regions from other vertebrate species. In this study, we indeed found the important pattern and most of the dinucleotide $(CA)_n$ were located in the highly conserved regions among various species. Consequently, these microsatellites can stand a good chance to dominate mammalian size variety, and we will conduct further experiments to decipher the size variation enigma.

2 System Description

In the developed system for retrieving microsatellites, there are ten representative species collected for comparative genomics analysis. A database was created which facilitates the search for microsatellites and provides the information of corresponding primers for PCR from different species. However, performing *in silico* analysis of biological data sometimes attempts to result in very high false positive rates. In order to promote the specificity of discovering important microsatellites form the proposed system, we take advantage of evolutionarily conserved segments among sequences from various species. Users are able to choose specific species or specific genes as comparing targets, through an orthologs look-up table, the system will filter out microsatellites which are not located in conserved regions. Screening processes narrow down candidate microsatellites and improve the performance of specificity of characteristics in functional gene annotation. Taking the IGF1 gene as the target, Alleles of a dinucleotide $(CA)_n$ and $(TG)_n$ microsatellites are found in the promoter regions from several vertebrate species, and the relative positions are displayed in the resulting figures.

[1]Department of Computer Science and Engineering, National Taiwan Ocean University, Email: twp@mail.ntou.edu.tw
[2]Institute of Bioscience and BioTechnology, National Taiwan Ocean University.
[3]Department of Aquaculture, National Taiwan Ocean University. Center for Marine Bioscience and Biotechnology, No. 2, Pei-Ning Rd, Keelung, Taiwan 20224, Republic of China.

3 Examples

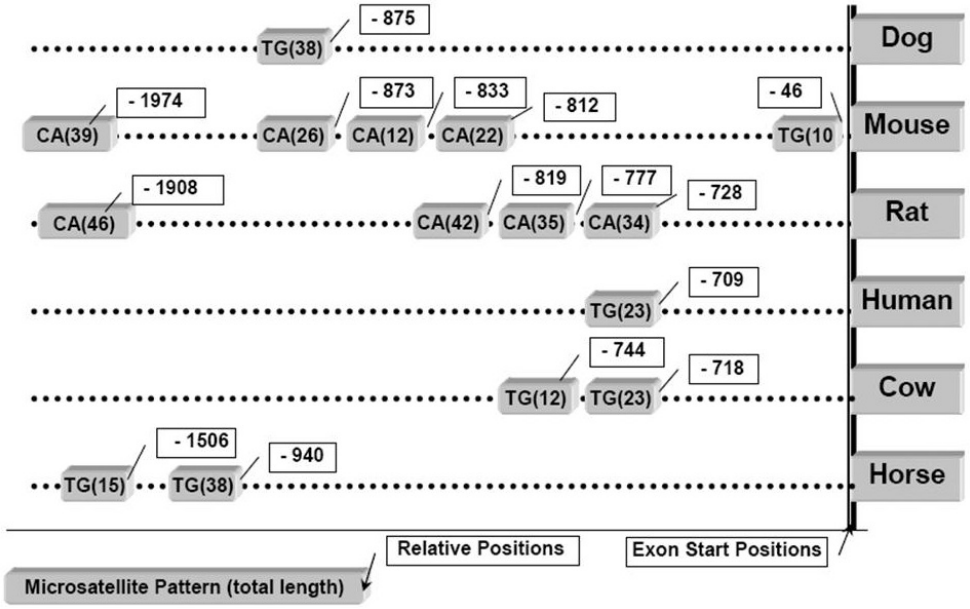


Figure 1: Alleles of a dinucleotide $(CA)_n$ and $(TG)_n$ microsatellites found in promoter regions of IGF1 gene in several vertebrate species.

References

[1] Li YC, et al. Microsatellites within genes: Structure, function, and evolution. *Molecular Biology and Evolution*, 21(6):991–1007, 2004.

[2] Kashi Y, King DG. Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics*, 22(5):253–259, 2006.

[3] Sutter NB, et al. A single IGF1 allele is a major determinant of small size in dogs. *Science*, 316:112–115, 2007.

Investigating the Promise of Extrinsic Similarity Measures for Gene Expression Analysis

Duygu Ucar,[1] Fatih Altiparmak,[1] Hakan Ferhatosmanoglu,[1] Srinivasan Parthasarathy[1,2]

1 Introduction

Due to advances in microrray technology, a wealth of information on the expression of genes during the life cycle of an organism is accumulated. To analyze and mine these datasets for vital information, various techniques and ideas have been proposed. Of particular interest to many scientists is the problem of identifying gene groups that have similar expression patterns over various samples, known as co-expressed genes. Genes with similar cellular functions have been theorized to behave similarly over different conditions. Thus, an effective similarity measure is essential to draw valuable conclusions from gene expression studies. A prevailing technique is to calculate the similarity of two genes based on their expression levels over all samples. In the follow-up studies, these pairwise similarities are accumulated in the form of interaction networks where genes are denoted as the nodes and two nodes are linked if the corresponding genes are significantly correlated across the samples.However, given the noise inherent in these datasets, these measure may not be adequate to distinguish random gene pairs from those that react similarly to changing conditions. We argue that since any given gene is likely to fluctuate in its measured expression level due to many possible sources of error, a similarity based on measurements of two genes (i.e., intrinsic) is more error-prone than a similarity based on relative positions of these two genes with respect to many genes (i.e., extrinsic). In addition, inferring the similarity of two genes based on their relations with a set of other genes will be in accordance with the biological hypothesis about gene products acting as complexes to accomplish certain cellular level tasks. Here, we will investigate an extrinsic way of deducing gene similarity from microrray studies and demonsrate the efficacy of extrinsic measures in inferring pairwise gene similarity, in constructing gene networks and in clustering genes.

2 Similarity Measures

In a typical microarray experiment, each gene is expressed at some certain level at each condition which is defined as the expression profile of the gene. In the context of microarray analysis, the intrinsic similarity of two genes is defined on the expression profiles of these two genes, where the most commonly used measure for microarray analysis is the Pearson Correlation Coefficient. Recently, Ravasz et al. proposed the Topological Overlap Measure (TOM) which takes into a step in incorporating external information, i.e., number of common neighbors, to infer similarity of two nodes in a biochemical network [3].

On the other hand, extrinsic similarity of two attributes (i.e., genes) is defined over other attributes in the dataset [2]. Before defining the specifics of an extrinsic similarity measure, a general definition can be given as follows:

$$ES_P(i,j) = \sum_{k \in P} |f(i,k) - f(j,k)| \quad (1)$$

Here, $f(i,k)$ denotes a function that signifies association between attributes i and k. P refers to the set of attributes that will contribute to the extrinsic similarity of attributes i and j.

As noted by Das et al [2], proper choice of the *Attribute Set* P and *Association Function* f is crucial for the usefulness of the resulting *extrinsic* measure. To derive an efficient extrinsic measure for microarray analysis, we first identify a gene set, P, that will be used to infer the extrinsic similarity of two genes. We propose to include genes that are similar to both of the genes under question to this set. Thus, initially we identify a set of genes that are intrinsically similar to a gene (say i), named as the neighborhood list of a gene (say N_i). Next, the *Attribute Set* of two genes (say i and j) is designated as the intersection

[1]Department of Computer Science and Engineering, The Ohio State University, OH, USA.
[2]To whom correspondence should be addressed. Email: srini@cse.ohio-state.edu

of their neighborhood lists (i.e., $P_{ij} = N_i \cap N_j$). Next, we determine which *Association Function* to use in our extrinsic similarity calculations. Das et al [2], proposed using the confidence of association rules in an application on market basket dataset, which we apply on our problem and compare against our own measures. However, we propose to base our *Association Functions* to the co-occurrence patterns of gene pairs in neighborhood lists, which we refer as Mutual Independence of genes in the rest of this section. Accordingly, we explore three possible type of *Co-occurrence Relations* between any two genes: co-occurring, independent or non-co-ocurrring. We used two existing independence tests to leverage *Co-occurrence Relations* between genes: Specific Mutual Information and a signed version of Chi-Square Independence Test. Our hypothesis is that if two genes have similar *Co-occurrence Relations* with the same set of genes, then they are extrinsically similar to each other.

3 Experiments

We evaluate the efficacy of extrinsic measures on a well-studied cancer dataset, which is composed of gene expression values of 62 colon tissue samples where the Affymetrix Hum6000 array with 6819 probes is used [1].

In our first experiment, we compare gene pairs that are labeled as 'similar' according to the discussed measures. For each measure, gene pairs are sorted starting from the most 'similar' one. We calculated semantic similarity of all the annotated pairs and calculate the average semantic similarity in each case [4]. This measure quantifies biological relevance of two genes with respect to the significance of their shared Gene Ontology (GO) annotations. As can be seen in Figure 1(a), the pairs identified with the *SMI*, *Chi* and *Confidence measures* show greater biological relevance when compared to the pairs identified by the other measures. For the top 1000 pairs, the improvement in the average semantic similarity score is up to 18%. The improvement obtained by using TOM measure is not as significant as that of the extrinsic measures.

Next, we construct association networks by connecting the top scoring gene pairs identified by each measure. Here, nodes represent genes, and two nodes are linked if the corresponding genes are 'similar' to each other. To keep the same size for all networks, we only used the top 0.01% of 'similar' gene pairs in each case. We calculate the average semantic similarity of pairs (i.e., edges) in each network and observe that *Chi* network has a score of 1.48, whereas the Pearson network only has 1.41. We also examine the quality of clusters extracted from these networks. To identify dense regions from our networks, we employ the most commonly used clustering algorithm, i.e., hierarchical clustering with average linkage. Each network is partitioned into 200 clusters, and each clustering arrangement is validated using the enrichment score, i.e., p-value, that signifies the statistical value of the functional homogeneity of a cluster. The p-value distributions for the significant clusters extracted from various gene association networks are shown in Figure 1(b). As can be observed from the figure, extrinsic similarity measures produce more number of clusters that are functionally homogenous.

Our experimental results prove that using the extrinsic measures, it is possible to identify gene pairs that are biologically more relevant. In addition, association networks generated based on these measures are shown to be more informative and useful for further analysis.

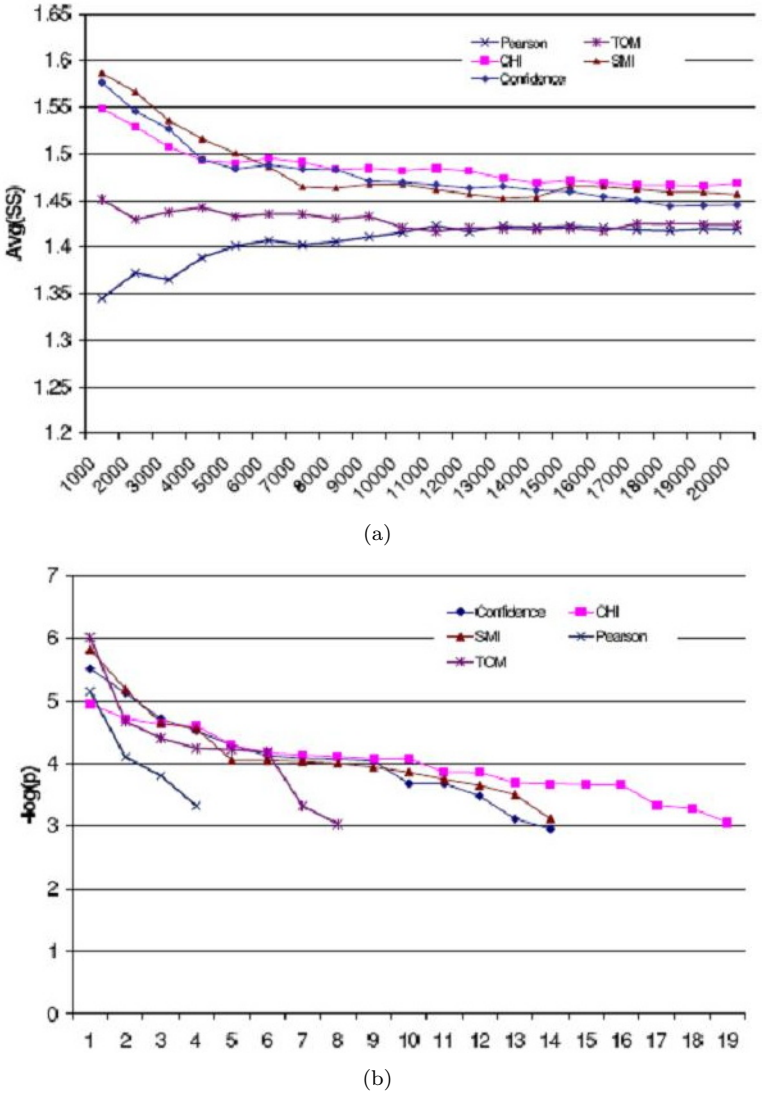
References

[1] U. Alon, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. USA*, 96:6745–6750, 1999.

[2] G. Das, H. Mannila, P. Ronkainen. Similarity of attributes by external probes. In: *Proc. 4th International Conference on Knowledge Discovery and Data Mining* (KDD-98), pages 23–29, 1998.

[3] E. Ravasz, et al. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[4] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In: *Proc. 14th International Joint Conference on Artificial Intelligence*, 1:448–453, 1995.



(a)

(b)

Figure 1: (a) Average Semantic Similarity of most similar pairs. (b) Enrichment score of clusters.

A Dense Graph Model for Haplotype Inference

Sharlee Climer,[1] Alan R. Templeton,[2] Weixiong Zhang[1]

Phasing genotype data to identify the composite haplotype pairs is a widely-studied problem due to its value for genome-wide association studies, population genetics research, and other significant endeavors. The accuracy of the phasing is crucial as identification of haplotypes is frequently the first step of expensive and vitally important studies. We present a combinatorial approach to this problem, which we call Splitting Heirs, that is based on a dense graph model. We have tested Splitting Heirs with several popular existing phasing methods including PHASE, HAP, and Pure Parsimony, on seven sets of real biological haplotype data. Our method yields the highest accuracy obtainable by these methods in all cases. Furthermore, Splitting Heirs is robust and had higher accuracy than any of the other approaches for the two data sets with high recombination rates. The success of Splitting Heirs validates the assumptions made by the dense graph model and highlights the benefits of finding globally optimal solutions.

Many methods used for haplotype inference have favored reduction of the cardinality of unique haplotypes. Pure Parsimony [2] is an extreme case in which a set of haplotypes are found such that the number of unique haplotypes is the least possible. Splitting Heirs favors reduced cardinality, but simultaneously considers other favorable properties and does not always yield a strictly parsimonious solution.

Some of the previous algorithms (e.g. PHASE [6]) have favored additional properties, such as pair-wise similarities between haplotypes. That is, if a potential haplotype is similar to one in the current solution, it is favored. In contrast, Splitting Heirs favors *cluster-wide* similarities by favoring solutions in which many haplotypes are similar to a number of other haplotypes. The dense graph model can be used to quantify the quality of a solution with regard to reduced cardinality and cluster-wide similarities.

Let h equal the number of unique haplotypes in a solution. Consider a graph with h nodes, in which each node represents a haplotype in the solution. The weight on an edge in the graph is set equal to the distance between the two haplotypes that are endpoints of the edge. Distances between haplotypes can be defined in various ways. A simple distance measure is just the number of sites in which they differ. If pair-wise similarities were the only concern, a graph to consider would contain only edges that connect each haplotype with its nearest neighbor. When relying completely on simple pair-wise distances, it is possible to have $h/2$ disjoint subgraphs with arbitrarily large distances between them. In real populations, we would expect to find clusters of haplotypes that are similar to each other, so it is desirable to enforce similarities beyond a single nearest neighbor.

In a *dense* graph model, the density of the graph is required to be greater than or equal to a given value, α. The density of a graph can be defined as e/h, where e is the number of edges in the graph. By considering these additional edges, similarities beyond single nearest neighbors are taken into consideration. We evaluate the quality of the dense graph solution using:

$$C_D = \sum_{i=1}^{e} w_i d_i + \sum_{i=1}^{h} u_i \tag{1}$$

where d_i is the distance of edge i and w_i and u_i are weights. In our experiments, we used a constant u_i value and $w_i = 1$ for all i.

The dense graph with the minimum cost C_D is considered optimal. We have cast this model as an Integer Linear Program (IP). The constraints of our IP require that the selected haplotypes resolve all of the genotypes. These constraints are similar to the constraints for the Pure Parsimony IP formulation. The key differences between our IP and the Pure Parsimony IP is that our objective function is Equation (1), and we add the following constraint to ensure the density of the graph: $e/h \geq \alpha$. Like Pure Parsimony, this problem may require exponential time to compute in the worst case. However, we were able to obtain globally optimal solutions using ILOGs Cplex 8.11, which is a generic IP solver.

On some occasions, the optimal dense graph may have more than one pair of haplotypes that can resolve a given genotype. When this is the case, Splitting Heirs assumes that common haplotypes are very common, and assigns the pair that contains the haplotype with the highest frequency in the set. Alternate pairs, along with their frequencies, are also provided for the user.

[1]Dept of Computer Science and Engineering, Washington University in St. Louis, MO, USA. Email: sharlee@climer.us, zhang@cse.wustl.edu

[2]Dept of Biology, Washington University in St. Louis, MO, USA. Email: temple_a@biology.wustl.edu

The dense graph model is biologically intuitive as it utilizes three widely accepted principles: the number of unique haplotypes within a given population is relatively small, many haplotypes are similar to others, and common haplotypes are very common. PHASE incorporates the first two of these principles in its priors. However, PHASE favors pairs of haplotypes that are similar. It is biologically intuitive that clusters of haplotypes are similar, not just pairs. Splitting Heirs effectively incorporates this intuition.

We have tested the biological accuracy of various haplotype inferencemethods using seven sets of true haplotype data derived experimentally (i.e. the individual haplotypes were identified, not the melded pairs). The first source of data used for comparisons is a set of 80 human *ApoE* haplotype pairs, each with nine SNPs [5]. These SNPs are drawn from the *apolipoprotein E* locus. Data set A in Table 1 is composed of these 80 pairs of haplotypes.

The second source [1] contains 39 pairs of human haplotypes, each with 411 sites, in a 48 kb region containing the *KLK13* and *KLK14* genes. There is a substantial amount of missing data in this set. Pure Parsimony, EM-DeCODER, and the current implementation of Splitting Heirs all require complete data. Six regions of complete data from this set are used for this study and correspond to data sets B through G in Table 1. They range from 5 sites to 47 sites in length. The 17 sites of set D have no recombination and are combined with 9 additional sites, which have a low recombination rate, to make set F.

We compare Splitting Heirs with several popular haplotype inference methods: Pure Parsimony [2], HAP [3], EMDeCODER [4], and PHASE [6]. Two of these implementations use combinatorial methods and the other two use statistical approaches. Table 1(a) shows the results for the data sets with little or no recombination. These results list the number of genotypes incorrectly phased as well as the total number of sites that were incorrectly phased by each method. As shown in the table, Splitting Heirs did better than, or as well as, all of the other solvers in every case. Table 1(b) contains the results for data with high recombination rates. Splitting Heirs outperformed the other algorithms on both data sets.

Due to consequences for vitally important genome-wide association studies and population genetics studies, the benefits of accuracy for the haplotype inference problem cannot be measured by mere financial gains. Splitting Heirs finds globally optimal solutions for this problem that favor low cardinality of unique haplotypes as well as similarities across clusters of haplotypes. Favoring cluster-wide similarities is biologically intuitive and this assumption is experimentally validated using true haplotype data.

References

[1] A.M. Andrés, A. G. Clark, E. Boerwinkle, et al. Assessing the accuracy of statistical haplotype inference with sequence data of known phase. *Genet. Epi.*, 31:659–671, 2007.

[2] D. Gusfield. Haplotype inference by pure parsimony. In: *14th Annual Symposium on Combinatorial Pattern Matching* (CPM03), pages 144–155, 2003.

[3] E. Halperin, E. Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20:1842–1849, 2004.

[4] T. Niu, Z. Qin, X. Xu, J. Liu. Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. *American Journal of Human Genetics*, 70:157–169, 2002.

[5] S. H. Orzack, D. Gusfield, J. Olson, et al. Analysis and exploration of the use of rule-based algorithms and consensus methods for the inferral of haplotypes. *Genetics*, 165:915–928, 2003.

[6] M. Stephens, N. Smith, P. Donnelly. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68:978–989, 2001.

			# of Gen. Wrong					# Het.	# of Sites Wrong				
	Data	n	Pars	HAP	EM	PHASE	Split	Sites	Pars	HAP	EM	PHASE	Split
(a)	A	80	5	11	10	4	4	154	5	13	13	4	4
	B	39	4	4	4	4	4	44	4	4	4	4	4
	C	39	1	0	0	2	0	136	3	0	0	3	0
	D	39	1	1	2	0	0	52	1	2	3	0	0
	E	39	3	3	-	4	1	193	7	3	-	6	1
(b)	F	39	4	11	-	6	4	186	5	12	-	8	4
	G	39	14	20	-	7	5	257	27	44	-	8	6

Table 1: Results for genotype sets with (a) little or no recombination and (b) high recombination rates. Number of genotypes in the set (n); the number of genotypes incorrectly phased by Pure Parsimony (Pars), HAP, EM-DeCODER (EM), PHASE, and Splitting Heirs (Split); total number of heterozygous sites in set of genotypes; and number of sites incorrectly phased by each method are tabulated.

Large-Scale Inference of Condition-Specific Regulation Using Gene Expression Data and Predicted Transcription Factor Occupancy of Promoters

Neil D. Clarke,[1] Hock Chuan Yeo, Zhen Xuan Yeo, Ye Li

Gene expression experiments have been performed under many different conditions. In contrast, large-scale ChIP-chip experiments (i.e., those involving many transcription factors) have been performed under just a few. It is likely, therefore, that only a fraction of condition-specific functional binding sites have been identified. Computational methods are required to further correlate factors, conditions, and target genes in order to infer more comprehensive regulatory networks, and to generate hypotheses that can be tested by directed ChIP experiments.

We have previously developed a method for predicting the probability of transcription factor binding to a promoter [1]. The method models cooperative and competitive binding in a physically meaningful manner, and appropriately uses protein concentration as a parameter. Genomewide nucleosome location data has also been incorporated into the model to improve the prediction of bound sites [2]. We are now using this method to systematically compare predicted binding profiles for over a hundred transcription factors to the changes in gene expression in hundreds of microarray experiments, and have identified many conditions under which predicted binding is significantly correlated with gene regulation. A joint probability analysis, using gene expression changes and predicted binding probabilities, further identifies the genes that are most likely to be direct targets of the transcription factor under that condition. This analysis recapitulates interactions inferred from expression and ChIP-chip analyses, and makes novel predictions that can be tested by ChIP experiments under previously unexplored conditions.

References

[1] Granek JA, Clarke ND. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol*, 2005, 6:R87.

[2] Liu X, Lee CK, Granek JA, Clarke ND, Lieb JD. Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome Res*, 2006, 16:1517–1528.

[1] Genome Institute of Singapore.

Organizers

Supporting Organizers

Platinum Sponsors

Gold Sponsor

Silver Sponsor