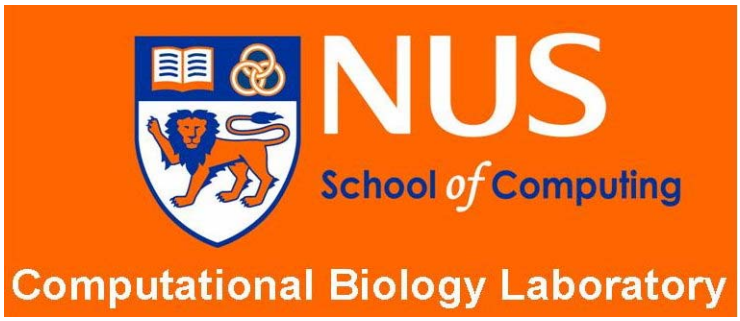


1 April 2005 to 31 March 2007

The frontier of medical science has rarely been as exciting and as full of opportunity as it is today. From basic science to clinical research to health services research, the opportunities made available through the impressive recent advances in the biomedical, physical, and computational sciences have brought us to a place of unprecedented opportunity. It is also now widely appreciated that present-day biomedical researchers are confronted by vast amounts of data from genome sequencing, microscopy, high-throughput analytical techniques for DNA, RNA, and proteins, and a host of other new experimental technologies. Coupled with enormous advances in computing power, it is expected that this flow of information should enable scientists to model and understand biological systems in novel ways.

Over the last decade, however, the community has come to the realization that, to make effective use of the explosion of data, new challenges must be faced especially in the development and application of computational analysis and knowledge discovery technologies and methodologies. The SOC Computational Biology Lab was thus established. The research of our lab will benefit biological and clinical researchers by enabling extraction and inference of new knowledge and value-added information from diverse biomedical data. Our research projects and activities are described below.

Protein Structure Comparison and Database Search

Background

Analysis of 3-dimensional (3D) protein structures plays an important role in bioinformatics. Since the functions of a protein is more closely related to its 3D structure than to its amino acid sequence, the study of proteins from structural perspective can give us more valuable information about their functions.

Objectives

There are two main objectives: (a) develop efficient and effective methods to compare a pair of 3D protein structures; (b) develop efficient and effective methods to search a database of 3D protein structures. The “efficiency” aspect deals with the speed of retrieval. The “effectiveness” aspect deals with the accuracy of the methods.

Achievements

We have designed a fine-grained protein structure alignment method named MatAlign. It is a two-step algorithm. First, we represent 3D protein structures as 2D distance matrices, and align these matrices by means of dynamic programming in order to find the initially aligned residue pairs. Second, we refine the initial alignment iteratively into the optimal one according to an objective scoring function. We have implemented MatAlign and evaluated it against two widely used structural comparison tools, DALI and CE. On the benchmark set of 68 protein structure pairs by Fischer et al., MatAlign provides better alignment scores, according to four different criteria, than both DALI and CE in a majority of cases. MatAlign is about 4 times faster than DALI, and has about the same speed as CE. The current implementation runs on Win32, and the executables can be downloaded from http://www1.i2r.a-star.edu.sg/~azeyar/genesis/MatAlign/MatAlign_v11_Windows.zip.

We have also proposed a rapid protein structure database retrieval system named ProtDex2, in which we adopt the techniques used in information retrieval (IR) systems in order to perform rapid database searching without having to access every structure in the database. The retrieval process is based on the inverted index constructed on the feature vectors of the relationships between the secondary structure elements (SSEs) of all the protein structures in the database. Experimental results show that ProtDex2 is very much faster than two commonly used detailed structural comparison methods, DALI and CE, yet not much sacrificing on the accuracy of the comparison. When comparing with a fast database scan method, Topscan, ProtDex2 is much faster and still slightly more accurate. The software is available at http://www1.i2r.a-star.edu.sg/~azeyar/genesis/ProtDex2.

Relevant Publications

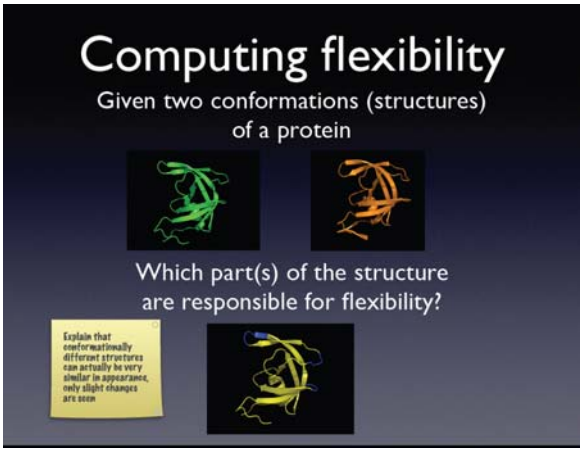
1. Z. Aung, K.-L. Tan. **MatAlign: Precise Protein Structure Comparison by Matrix Alignment**. *Journal of Bioinformatics and Computational Biology*, 4(6):1197--1216, December 2006.
2. Z. Aung and K.L. Tan. **Automatic 3D Protein Structure Classification Without Structural Alignment**. *Journal of Computational Biology*, 12(9):1221--1241, November 2005.
3. C.H. Chionh, Z. Huang, K.L. Tan, Z. Yao. **Towards Scaleable Protein Structure Comparison and Database Search**. *International Journal on Artificial Intelligence Tools*, 14(5):827--848, October 2005.

Protein Structure and Protein Motion

Background

Protein conformational flexibility plays a critical role in vital biological functions, such as immune protection and enzymatic catalysis. An example is the “flap” motion of HIV protease, a major inhibitory drug target for AIDS therapy. The flaps, located near the reactive site of HIV protease, must open to allow a ligand to bind and then close to form direct contacts with the ligand. Such motion is an essential means which proteins rely on to perform their functions, and therefore provides an important link between structure and function, a central relationship in molecular biology.

Objectives



We envision that future protein databases will contain not only static structures of proteins, as the Protein Data Bank (PDB) does currently, but also the motion of proteins. Users can submit queries on protein motions to study the relationship between structure and function, just as they submit queries on sequences and static structures today. Such information will greatly enhance our ability to understand and predict ligand-protein and protein-protein interactions during processes such as pharmaceutical drug design. Towards this goal, we are developing geometric computation tools to explore, analyze, and model protein conformational flexibility.

Achievements

We have developed a new method called Stochastic Roadmap Simulation (SRS) for studying protein folding kinetics and have used it to predict experimental quantities in protein folding computationally. Interesting properties of molecular motion are best characterized statistically by considering an ensemble of motion pathways rather than an individual one. Classic simulation techniques, such as the Monte Carlo method and molecular dynamics, generate individual pathways one at a time and are easily "trapped" in the local minima of the energy landscape. They are computationally inefficient if applied in a brute-force fashion to deal with many pathways. SRS uses a randomized technique for sampling molecular motion and exploring the kinetics of such motion by examining multiple pathways simultaneously.

We have developed an algorithm and related software called pFlexAna for detection of protein conformational changes from experimental data obtained through, e.g., X-ray crystallography. Due to noise in data, determining salient conformational changes accurately and efficiently is a challenging problem. A key element of pFlexAna is a statistical test that determines the similarity of two protein structures in the presence of noise. Using data from the Protein Data Bank and the Macromolecular Movements Database, we tested the algorithm on proteins that exhibit a range of different conformational changes. Results show that our algorithm can reliably detect salient conformational changes, including well-known examples such as hinge and shear.

With collaborators from the NUS School of Medicine, we are studying the protein-protein interaction between Bcl-2 and Rac1.

Relevant Publications:

1. A. Nigham, D. Hsu. **Protein Conformational Flexibility Analysis with Noisy Data**. *Proceedings of 11th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 396--411, San Francisco, CA, April 2007.

2. T.-H. Chiang, M. S. Apaydin, D. L. Brutlag, D. Hsu, and J.-C. Latombe. **Predicting Experimental Quantities in Protein Folding Kinetics using Stochastic Roadmap Simulation**. *Proceedings of 10th Annual International Conference on Research in Computational Molecular Bioloy (RECOMB)*, pages 410--424, Venice, Italy, April 2006.

Increasing Reliability of Protein Interactomes

Background

Progress in high-throughput experimental techniques in the past decade has resulted in a rapid accumulation of protein-protein interaction (PPI) data. However, recent surveys reveal that interaction data obtained by the popular high-throughput assays such as yeast-two-hybrid experiments may contain as much as 50% false positives and false negatives.

Reliability of Protein Interactome: Are We There Yet?

	Coverage	Data quality
DNA genome sequence	99% of genome sequence	99.9% correct
mRNA profiling	80-90% of transcripts represented	90% of spots are good data
Protein interaction data	10-30% of interactions catalogued	50-70% of interactions are spurious

As a result, further carefully-focused small-scale experiments are often needed to complement the large-scale methods to validate the detected interactions. However, the vast interactomes require much more scalable and inexpensive approaches.

Objectives

There are three main objectives: (a) identify properties that characterize true-positive and false-positive PPIs; (b) develop efficient and effective methods for assessing the reliability of PPIs reported in high-throughput assays; (c) develop efficient and effective methods for identifying false-negative PPIs and new protein complexes from high-throughput assays.

Achievements

We have developed several methods for assessing the reliability of a PPI, given a graph derived from high-throughput protein interaction experiments. The pairs of “interacting” proteins that are ranked highly by these methods are shown more likely to be true-positive interacting pairs. The most interesting feature of these methods is that they are able to rank the reliability of an interaction between a pair of proteins using only the topology of the interactions between that pair of proteins and their neighbours within a short “radius”.

Our methods can be roughly divided into two groups. The first group is represented by the “functional similarity weighting” (FS-Weight) and “CD Distance” indices. This group of indices attempt to provide abstract mathematical characterizations of networks of reliable protein-protein interactions. They are based on the hypothesis that true-positive interactions are likely to be characterized by dense cross connections in the derived interaction graph. The second group is represented by the “meso-scale motifs”

(NeMoFinder) indices. This group of indices attempt to provide explicit motifs of network connections that are associated with reliable protein interactions.

The NeMoFinder, CD Distance, and FS-Weight are far superior (30-50% better correlated with function homologeneity, localization coherence, and gene expression correlation) than previous methods (e.g., interaction generality index) for detecting false positives in protein interaction experiments. CD Distance and FS-Weight are also very fast to compute. Furthermore, with small modifications such as incorporation of interacting motif pairs, these indices can also be used for false negative detection at high confidence.

Relevant Publications

1. J. Chen, W. Hsu, M. L. Lee, S.-K. Ng. **Increasing Confidence of Protein Interactomes Using Network Topological Metrics**. *Bioinformatics*, 22:1998--2004, 2006.
2. J. Chen, W. Hsu, M. L. Lee, S.-K. Ng. **NeMoFinder: Dissecting Genome-Wide Protein-Protein Interactions with Repeated and Unique Network Motifs**. *Proc. 12th ACM SIGKDD Interactional Conference on Knowledge Discovery and Data Mining (KDD)*, pages 106--115, Philadelphia, PA, August 2006.
3. J. Chen, H. N. Chua, W. Hsu, M.-L. Lee, S.-K. Ng, R. Saito, W.-K. Sung, L. Wong. **Increasing Confidence of Protein-Protein Inteactomes**. *Proceedings of 17th International Conference on Genome Informatics (GIW)*, pages 284--297, Yokohama, Japan, 18-20 December 2006. (invited keynote paper)
4. J. Chen, W. Hsu, M.L. Lee, and S.-K. Ng. **Discovering Reliable Protein Interactions from High-Throughput Experimental Data Using Network Topology**. *Artificial Intelligence in Medicine*, 35:37--47, 2005.

Protein Function Prediction

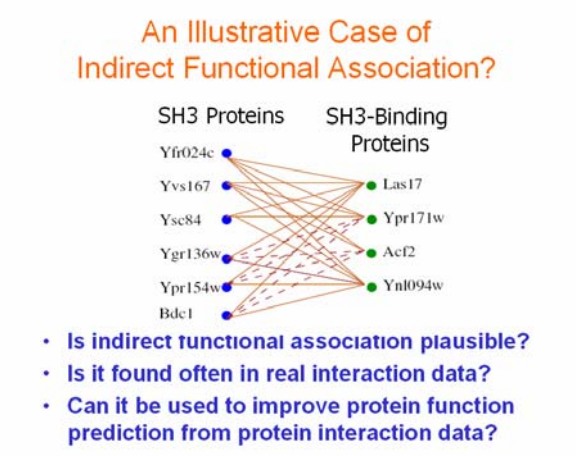
Background

Although sequence similarity search has been proven useful for protein function prediction in many cases, it has fundamental limitations. First, only a fraction of newly discovered sequences have identifiable homologous genes in the current databases. Second, the most prominent vertebrate organisms in GenBank have only a fraction of their genomes present in finished sequences. New bioinformatics methods should allow inference of protein function using “associative analysis” of functional properties to complement the traditional sequence homology-based methods. Associative properties that have been used to infer function not evident from sequence homology include: co-occurrence of proteins in operons or genome context; proteins sharing common domains in fusion proteins; proteins in the same pathway; proteins with correlated gene expression patterns; etc.

Objectives

In this project, we investigate and develop methods for inferring protein functions without sequence homology. In particular, (a) we find out how significant functional association between level-2 neighbors is; (b) we investigate how they can be exploited

for protein function prediction in a graph-based framework; and (c) we investigate how to integrate protein interaction information with other types of information to improve the sensitivity and specificity of protein function prediction, in a graph-based framework. At the end of the project, we expect to have developed a robust and powerful system to predict protein functions, even in the absence of sequence homology.



Achievements

We have developed FS-Weight, a method for assigning function to a protein based on the functions of direct and indirect interaction partners of that protein. The FS-Weight method predicts the functions of a protein in two steps: (1) assigning a weight (the FW-Weight) to each of its level-1 and level-2 neighbours by estimating its functional similarity with the protein using the local topology of the interaction network as well as the reliability of experimental sources, and (2) scoring each function based on its weighted frequency in these neighbours. FS-Weight makes predictions with better precision and recall compared to other protein interaction based methods (e.g., Neighbour Counting and Chi-Square) for the seven model genomes we have tested. FS-Weight is also robust in the presence of noise in the protein interaction data, maintaining consistent prediction performance even when a large number of interactions are randomly added or deleted to the interaction data.

We have also proposed a method called LaMoFinder to annotate network motifs with biological information associated with the proteins in a protein-protein interaction network. Our method is specifically devised to handle the large labeling space as well as the sophisticated Gene Ontology scheme in which the proteins were annotated. As a result, we have captured not only the topological shapes of the motifs, but also the biological context in which they occurred in the labeled network motifs. We also demonstrated how the network motifs labeled by LaMoFinder can be used to predict the functions of unknown proteins in the PPI network. Our superior performance against other current prediction methods confirmed that the network motifs have indeed been adequately enriched by LaMoFinder for the more sophisticated biological applications such as protein function prediction.

We have also developed a technique for identifying potential annotation errors in protein sequence databases. These databases are crucial to protein function prediction that uses the principle of “guilt by association” of sequence similarity. However, these databases are known to contain many annotation errors. We have derived a statistical method to filter and assess the reliability of data from these databases. Our experiments show that we can effectively detect mis-annotated sequence data.

Relevant publications

1. J. Chen, W. Hsu, M. L. Lee, S.-K. Ng. **Labeling network motifs in protein interactomes for protein function prediction**. *Proceedings of 23rd International Conference on Data Engineering (ICDE)*, pages 546--555, Istanbul, Turkey, April 2007.
2. H. N. Chua, W.-K. Sung, L. Wong. **Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein-Protein Interactions**. *Bioinformatics*, 22:1623--1630, 2006.
3. S.-H. Tan, W. Hugo, W.-K. Sung, S.-K. Ng. **A Correlated Motif Approach for Finding Short Linear Motifs from Protein Interaction Networks**. *BMC Bioinformatics*, 7:502, 2006.
4. K. Ning, H. N. Chua. **Automated Identification of Protein Classification and Detection of Annotation Errors in Protein Databases Using Statistical Approaches**. *LNBI 3886: Proceedings of PAKDD 2006 Workshop on Knowledge Discovery in Life Science Literature (KDLL2006)*, pages 123--138, Singapore, April 2006.

Identifying Functional Elements in Human and Other Genomes

Background

Interactions between macromolecules play many essential roles---e.g., metabolic reactions and signal transduction---and occur in many combinations, such as protein-protein, protein-DNA, and protein-RNA. Protein interactions with DNA and RNA are the primary mechanisms for controlling gene expression. What is needed is a recognition code that maps from the protein sequence to a pattern that describes the family of DNA binding sites---the functional elements. Identification of functional elements in the human genome is fundamental to our understanding of cell functions---how these codes orchestrate the complex network of gene transcription, the transcriptome, and interactions in distinct locations.

Objectives

(a) Develop methods for accurate identification of transcription factor binding sites and other regulatory sites. (b) Develop methods for inferring the interactions of transcription factors and other functional elements.

Achievements

We have developed an algorithm called LocalMotif to detect localized motifs in long regulatory sequences. A novel score function predicts the region of localization of the motif. This score is combined with other scoring measures including Z-score and relative entropy to detect the motif. The algorithm is optimized for fast processing of long regulatory sequences. Tests on simulated and real datasets confirm that LocalMotif accurately determines the region of localization of motifs and automatically discovers the biologically relevant motifs, which can be detected by other motif finding algorithms only when the search is restricted to the relevant interval.

Relevant Publications

1. V. Narang, W.-K. Sung, A. Mittal. **LocalMotif---An In Silico Tool for Detecting Localized Motifs in Regulatory Sequences**. *Proceedings 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2006)*, pages 791--799, Washington D.C., November 2006.
2. V. Narang, W.-K. Sung, A. Mittal. **Computational Annotation of Transcription Factor Binding Sites in *D. Melanogaster* Developmental Genes**. *Proceedings of 17th International Conference on Genome Informatics (GIW 2006)*, pages 14--24, Yokohama, Japan, December 2006.
3. H. L. Chan, T. W. Lam, W. K. Sung, Prudence W. H. Wong, S. M. Yiu, and X. Fan. **The Mutated Subsequence Problem and Locating Conserved Genes**. *Bioinformatics*, 21(10):2271--2278, 2005

Management and Analysis of Gene Expression Data

Background

The development of microarray technology has made possible the simultaneous monitoring of the expression of thousands of genes. This development offers great opportunities in advancing the diagnosis of diseases, the treatment of diseases, and the understanding of gene functions.

Objectives

Our main goals are: (a) develop techniques to derive gene regulatory networks from gene expression data; (b) develop technologies for the design of microarrays; (c) develop tools for optimization of disease treatment based on gene expression profiles.

Achievements

We have developed a new conditional dependence learning algorithm to learn a gene regulatory network from gene expression data. Besides the pairwise correlation of the gene expression profiles of a pair of genes, our algorithm considers three additional factors: (a) the collaboration among regulators, (b) the formation of regulatory complex, and (c) the variable time delay to learn the gene network. Experiments on both artificial and real-life gene expression datasets validate the effectiveness of the algorithm.

We have also proposed an efficient algorithm to identify time-lagged co-regulated gene clusters. The algorithm facilitates localized comparison and processes several genes simultaneously to generate detailed and complete time-lagged information for genes/gene clusters. Our results show that the algorithm is not only efficient, but also delivers more reliable and detailed information on time-lagged co-regulation between genes/gene clusters, compared to existing methods such as the event method and edge detection method.

Relevant Publications

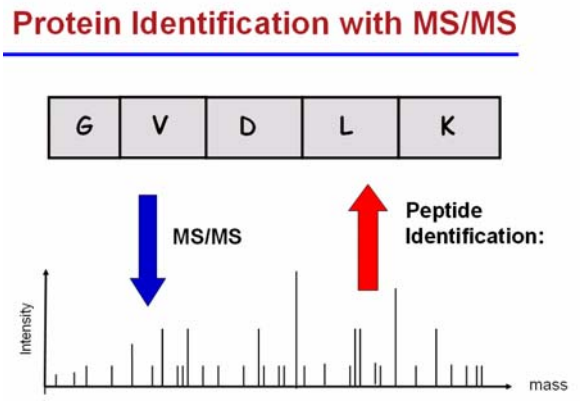
1. T.-F. Liu, W.-K. Sung, A. Mittal. **Model gene network by semi-fixed Bayesian network**. *Expert Syst. Appl.*, 30(1):42--49, 2006.

2. T.-F. Liu, W.-K. Sung, A. Mittal. **Learning Gene Network Using Time-delayed Bayesian Network**. *International Journal on Artificial Intelligence Tools*, 15(3):353--370, 2006.
3. K. I. Zeller, X.D. Zhao, C. W. H. Lee, K. P. Chiu, F. Yao, J. T. Yustein, H. S. Ooi, Y. L. Orlov, A. Shahab, H. C. Yong, Y.T. Fu, Z. Weng, V. A. Kuznetsov, W.-K. Sung, Y. Ruan, C. V. Dang, and C.-L. Wei. **Global mapping of c-Myc binding sites and target gene networks in human B cells**. *PNAS*, 103:17834--17839, 2006
4. L. Ji, K.-L. Tan. **Identifying time-lagged gene clusters using gene expression data**. *Bioinformatics*, 21(4):509--516, 2005.

Analyzing Mass Spectra for Identifying Proteins

Background

Proteomics is useful for understanding the expression of proteins in cells at different levels, at different time points, and in different forms. Such an understanding is critical to drug discovery and medical advances. Mass spectrometers are the predominant tool to accomplish some of the primary goals of proteomics. For example, identification of proteins, determination of expression level of proteins, and determination of post-translational modifications, sites, and types. Due to limitations in mass spectrometers and associated software tools, most of the MS/MS data generated by mass spectrometers are rejected because they are not interpretable by currently available software. Furthermore, the remaining data usually contain many false positives. In addition, the sensitivity and precision of mass spectrometers vary greatly.



Objectives

(a) Develop efficient, accurate, and robust algorithms for protein identification from tandem mass spectra. (b) Develop efficient, accurate, and robust algorithms for alignment of noisy mass spectra.

Achievements

Most peptide sequencing algorithms currently handle spectra of charge 1 or 2 and have not been designed to handle higher-charge spectra. We have recently given a characterization of multi-charge spectra by generalizing existing models and analyzed spectra with charges up to 5. Our analysis shows that higher charge peaks are present and they contribute significantly to prediction of the complete peptide. They also help to explain why existing algorithms do not perform well on multi-charge spectra.

We have further proposed an algorithm, GST-SPC, for de novo sequencing of multi-charge mass spectra. It computes a peptide sequence that is optimal with respect to shared peaks count from among all sequences that are derived from multi-charge strong tags. GST-SPC uses a larger set of multi-charge strong tags than existing methods, which

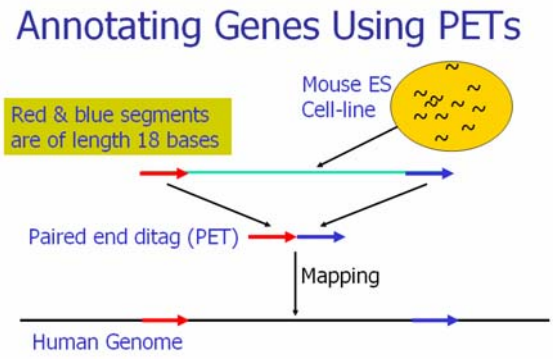
improves the theoretical upper bound on performance. Experimental results confirm the improvement.

Relevant Publications

1. K. Ning, K. F. Chong, H. W. Leong. **De Novo Sequencing for Mass Spectra based on Multi-Charge Strong Tags**. *Proceedings of 5th Asia-Pacific Bioinformatics Conference (APBC)*, pages 287--296, Hong Kong, January 2007.
2. K. F. Chong, K. Ning, H. W. Leong, P. A. Pevzner. **Modeling and Characterization of Multi-Charge Mass Spectra for Peptide Sequencing**. *Journal of Bioinformatics and Computational Biology*, 4(6):1329--1352, December 2006.
3. K. F. Chong, K. Ning, H. W. Leong, P. A. Pevzner. **Characterization of Multi-Charge Mass Spectra for Peptide Sequencing**. *Proceedings of 4th Asia-Pacific Bioinformatics Conference (APBC)*, pages 109--119, Taipei, Taiwan, February 2006.
4. K. Ning, K. F. Chong, H. W. Leong. **A Database Search Algorithm for Identification of Peptide with Multiple Charges Using Tandem Mass Spectrometry**. *LNBI 3916: Proceedings of PAKDD 2006 Workshop on Data Mining for Biomedical Applications (BioDM 2006)*, pages 2--13, Singapore, April 2006
5. K. Ning, H. K. Ng, H. W. Leong. **PepSOM: An Algorithm for Peptide Identification by Tandem Mass Spectrometry Based on SOM**. *Proceedings of 17th International Conference on Genome Informatics (GIW 2006)*, pages 194--205, Yokohama, Japan, December 2006.

Annotating the Genome Using Paired-End-diTag

Background



Annotating the whole genome is an expensive and time-consuming task. Recently, there is a new technology called Paired-End-diTag (PET). This technology allows us to find the 20 bps of the two ends of genes/transcripts in a high-throughput and cost-effective way. When coupled with chromatin immunoprecipitation (ChIP), we can also precisely extracting PETs surrounding all transcription factor binding sites in whole genome. Since PETs are short and the genome is big, the bioinformatics challenge is to have a method which can align the PETs onto the genome accurately and efficiently.

Objectives

Our main aim is to develop a tool to help biologists process the PET data. The tool should be accurate and efficient. In addition, cancer cell-lines may generate some abnormal genes known as fusion genes due to genome rearrangement in cancer cells. Our second aim is to discover these genes.

Achievements

We have developed a mapping algorithm that can efficiently and accurately align PETs onto the genome. This allows us to discover numerous novel genes and novel transcription factor binding sites in human and mouse genomes. In addition, we have discovered a number of fusion genes in cancer cells. To help biologists, our algorithm is implemented as a pipeline PET-Tool which allows the biologists to manage and process the PET sequence data.

Relevant Publications

1. K. P. Chiu, C.-H. Wong, Q. Chen, P. Ariyaratne, H. S. Ooi, C.-L. Wei, W.-K. Sung, Y. Ruan. **PET-Tool: A Software Suite for Comprehensive Processing and Managing of Paired-End diTag (PET) Sequence Data**. *BMC Bioinformatics*, 7:390, 2006.
2. P. Ng, J. Tan, H. S. Ooi, Y. L. Lee, K. P. Chiu, M. Fullwood, K. Srinivasan, C. Perbost, L. Du, W.-K. Sung, C.-L. Wei, Y. Ruan. **Multiplex Sequencing of Paired-End Ditags (MS-PET): A Strategy for the Ultra-High-Throughput Analysis of Transcriptomes and Genomes**. *Nucleic Acids Research*, 34(12):e84, 2006.
3. P. Ng, C.-L.Wei, W.-K. Sung, K. P. Chiu, L. Lipovich, C. C. Ang, S. Gupta, A. Shahab, A. Ridwan, C. H. Wong, E. T. Liu, Y. Ruan. **Gene Identification Signature (GIS) Analysis for Transcriptome Characterization and Genome Annotation**. *Nature Methods*, 2:105--111, 2005.

RNA Secondary Structure Prediction

Background

Due to advances in sequencing technologies, many RNA sequences have been discovered. Moreover, only a few of their structures have been deduced. On the other hand, the chemical and biological properties of many RNAs (like tRNAs) are determined primarily by their secondary structures. Therefore, determining the secondary structures of RNAs is becoming one of the most important topics in bioinformatics.

Objectives

We consider two problems related to RNA: The first problem is on inferring the RNA secondary structure of a RNA sequence by comparing it with another RNA sequence whose secondary structure is known. The second problem is on extracting local sub-structures shared by two RNA secondary structures. Note that such a sub-structure may be the functional part of the two RNAs.

Achievements

We expect two RNAs having similar function to generally have similar structure. We have devised a polynomial time dynamic programming algorithm to infer the secondary structure for a RNA sequence based on the known secondary structure of another functionally similar RNA.

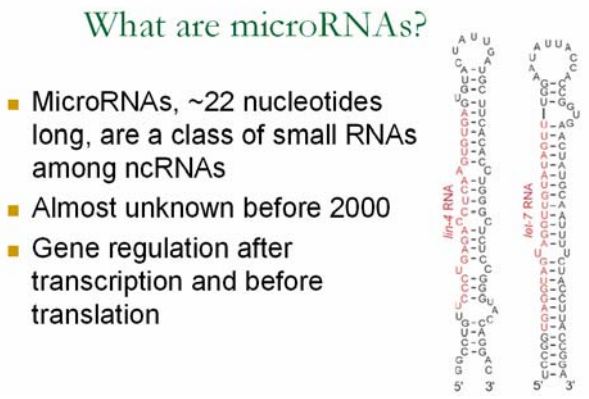
For the local sub-structure problem, we propose to model the local structural motif as a local gapped subforest. Our model is general enough to describe structural motif like SECIS-motif. We have further proposed a polynomial time algorithm to compute the local gapped subforest of two RNA secondary structures.

Relevant Publications

1. J. Jansson, N. T. Hieu, W.-K. Sung. **Local Gapped Subforest Alignment and Its Application in Finding RNA Structural Motifs** . *Journal of Computational Biology*, 13(3):702--718, April 2006.
2. J. Jansson, S.-K. Ng, W.-K. Sung, H. Willy. **A Faster and More Space-Efficient Algorithm for Inferring Arc-Annotations of RNA Sequences through Alignment**. *Algorithmica*, 46(2):223--245, 2006.

Recognition of MicroRNA Precursors and their Targets

Background



MicroRNAs (miRNAs) are small non-coding RNAs of about 22 nucleotides long. They play an important regulatory functions in eukaryotic gene expression through mRNA degradation or translation inhibition. The regulatory functions of miRNAs range from cell proliferation, fat metabolism, neuronal patterning in nematodes, neurological diseases, modulation of hematopoietic lineage differentiation in mammals, development, cell death, cancer, and control of leaf and flower development in plants.

Objectives

(a) Experimental miRNA identification is technically challenging and incomplete. This calls for the development of computational approaches to complement experimental approaches to miRNA gene identification. We propose in this project to investigate de novo miRNA precursor prediction methods. (b) Analyzing the binding of miRNAs to their mRNA target sites reveals that many different factors determine what constitutes a good fit. We intend to investigate these factors in detail and to construct decision models for predicting miRNA targets. (c) Finally, we would like to understand the role of miRNAs in a number of human diseases.

Achievements

We follow the “generation, feature selection, and feature integration” paradigm of constructing recognition models for genomics sequences. We have generated and identified features based on information in both primary sequence and secondary structure, and use these features to construct accurate decision models for the recognition of miRNA precursors.

Relevant Publications

1. L. H. Yang, W. Hsu, M. L. Lee, L. Wong. **SVM-based Identification of microRNA Precursors**. *Proceedings of 4th Asia-Pacific Bioinformatics Conference (APBC)*, pages 267--276, Taipei, Taiwan, February 2006.
2. Y. Zheng, W. Hsu, M. L. Lee, L. Wong. **Exploring Essential Attributes for Detecting MicroRNA Precursors from Background Sequences**. *LNBI 4316: Proceedings of 2006 VLDB Workshop on Data Mining in Bioinformatics*, pages 131--145, September 2006.
3. Y. Z., C. K. Kwoh. **Informative MicroRNA Expression Patterns for Cancer Classification**. *LNBI 3916: Proceedings of PAKDD 2006 Workshop on Data Mining for Biomedical Applications (BioDM 2006)*, pages 143--154, Singapore, April 2006.

Sequence Indexing and Database Search

Background

One of the important daily tasks of biologists is performing homology search. Currently, heuristic methods like BLAST are used for this task, whose running time is linear to the size of the database (for instance, human genome is of size 3 billion bases). As the size of typical sequence databases is increasing rapidly, the performance of these tools is going to deteriorate. Thus it is important to have solutions to improve the performance of database search.

Objectives

Our aim is to improve the performance of database search by utilizing sequence indexing techniques. We consider two directions. The first direction is to create compressed index. The second direction is to rely on disk-based indexing.

Achievements

For compressed indexing, we devoted our effort on compressed suffix array, which is a compressed index of biological sequences. We have developed a number of approximate string matching algorithms utilizing the compressed suffix array to speedup the database search. Our solution has shown advantages in some specialized homology searching problem in the biology domain. For disk-based indexing, we have developed CPS-tree, which is a compact partitioned suffix tree on disk. For exact pattern searching, the performance of CPS-tree is very good and the pattern searching time is independent of the genome length.

Relevant Publications

1. S.-S. Wong, W.-K. Sung, L. Wong. **CPS-tree: A Compact Partitioned Suffix Tree for Disk-Based Indexing on Large Genome Sequences**. *Proceedings of 23rd IEEE International Conference on Data Engineering*, pages 1350--1354, Istanbul, Turkey, April 2007.

2. T. N. D. Huynh, W.-K. Hon, T. W. Lam, W.-K. Sung. **Approximate string matching using compressed suffix arrays**. *Theoretical Computer Scence.*, 352(1-3):240--249, 2006.
3. H.-L. Chan, T.-W. Lam, W.-K. Sung, S.-L. Tam, S.-S. Wong. **Compressed Indexes for Approximate String Matching**. *Proc. 16th Annual European Symposium on Algorithms (ESA 2006)*, pages 208--219, September 2006.
4. H.-L. Chan, T. W. Lam, W.-K. Sung, S.-L. Tam, S.-S. Wong. **A Linear Size Index for Approximate Pattern Matching**. *Proc. 17th Annual Symposium on Combinatorial Pattern Matching (CPM 2006)*, pages 49--59, July 2006.

Tag SNP Selection and Association Study

Background

A SNP (Single Nucleotide Polymorphism) is a specific location in our genome where different people have different DNA bases. Roughly speaking, there are about millions of SNPs in human beings. They are the major differences among different individuals. Studying SNPs is helpful to understanding genetic diseases.

Objectives

There are a number of bioinformatics problem related to SNPs. We target two particular problems: (a) tag SNP selection problem and (b) association study. Since there are millions of SNPs, it is difficult to investigate all the SNPs. The tag SNP selection problem is the selection of a subset of informative SNPs to represent the major variations among the individuals. Given a set of SNPs for a set of individuals, an association study is the finding of a combination of SNPs which has the potential to cause a genetic disease.

Achievements

We have developed a method for tag SNP selection. Our method efficiently selects a minimum set of tag SNPs that maximizes the independence among the selected SNPs. We have also developed a method for association study. Unlike most of the previous works which perform single SNP association, our method considers the association of variable-size haplotypes. Through regularized regression analysis, we tackle the problem of multiple degrees of freedom in haplotype test. Our method can handle a large number of haplotypes in association analyses more efficiently and effectively than do currently available approaches.

Relevant Publications

1. Y. Li, W.-K. Sung, J. J. Liu. **Association Mapping via Regularized Regression Analysis of Single-Nucleotide–Polymorphism Haplotypes in Variable-Sized Sliding Windows**. *American Journal of Human Genetics*, 80(4):705--715, April 2007.
2. T.-F. Liu, W.-K. Sung, Y. Li, J.-J. Liu, A. Mittal, P.-L. Mao. **Effective Algorithms for Tag SNP Selection**. *Journal of Bioinformatics and Computational Biology*, 3(5):1089--1106, 2005.

Phylogenetic Tree and Network Reconstruction

Background

A phylogenetic tree, also known as a cladogram or a dendrogram, is a tree describing several life forms and their relations. Phylogenetic network is a generalization of phylogenetic tree. In addition to the ancestral-descendent mutational relationship, phylogenetic networks capture evolutionary events such as horizontal gene transfer and hybridization. Constructing a phylogenetic tree/network is helpful to understanding the history of life and to analyzing rapidly mutating viruses like HIV. Phylogenetic tree is also an important tool in the comparative genomic. Through comparing multiple species, it helps to predict protein structure, gene expression pattern, etc. Hence, it is important to have efficient and accurate tools for reconstructing phylogeny.

Objectives

Our aim is to develop methods for constructing phylogenetic tree and network. In particular, we are interested in approaches based on combining trees.

Achievements

We have developed accurate and efficient methods for constructing phylogenetic tree and network. For phylogenetic network, we give the first polynomial time algorithm for combining a dense set of triplets into a galled phylogenetic network. Also, we generalize the method to reconstruct a network by combining a set of phylogenetic trees. For phylogenetic tree, we study the supertree problem. We have a fixed-parameter polynomial time algorithm to construct a maximum agreement supertree of a set of phylogenetic trees.

Relevant Publications

1. Y.-J. He, T. N. D. Huynh, J. Jansson, W.-K. Sung. **Inferring Phylogenetic Relationships Avoiding Forbidden Rooted Triplets**. *Journal of Bioinformatics and Computational Biology*, 4(1):59--74, 2006.
2. J. Jansson, N. B. Nguyen, W.-K. Sung. **Algorithms for Combining Rooted Triplets into a Galled Phylogenetic Network**. *SIAM Journal of Computing*, 35(5):1098--1121, 2006.
3. J. Jansson, W.-K. Sung. **Inferring a Level-1 Phylogenetic Network from a Dense Set of Rooted Triplets**. *Theoretical Computing Science*, 363(1):60--68, 2006.
4. J. Jansson, J. H. K. Ng, K. Sadakane, W.-K. Sung. **Rooted Maximum Agreement Supertrees**. *Algorithmica*, 43(4):293--307, 2005.
5. Trinh N. D. Huynh, Jesper Jansson, Nguyen Bao Nguyen, and Wing-Kin Sung. **Constructing a Smallest Refining Galled Phylogenetic Network**. *Proceedings of 9th Annual International Conference on Research in Computational Molecular Bioloy (RECOMB)*, pages 265--280, Cambridge, MA, May 2005.

Developing Bio-Dynamical Systems and Analyzing Regulatory Behavior of Biological Pathways

Background

Mathematical modeling has been increasing recognized as a useful tool in the biomedical sciences. Biological models provide a coherent framework for interpreting data. They can uncover new phenomena to explore, identify key factors of a system, link different level of details, enable a formalization of intuitive understanding, screen unpromising hypotheses, predict variable inaccessible to measurement, and expand the range of questions that can meaningfully be asked.

Objectives

We propose to study bio-dynamical systems, which are mathematical models of the dynamic behaviors of biological systems. Specifically, we will develop efficient computational representations, analysis techniques, and algorithms for modeling and reasoning about bio-dynamical systems. Our goal is to provide computational tools that enable biologists to design experiments more effectively and thus accelerate the process of biological discovery.

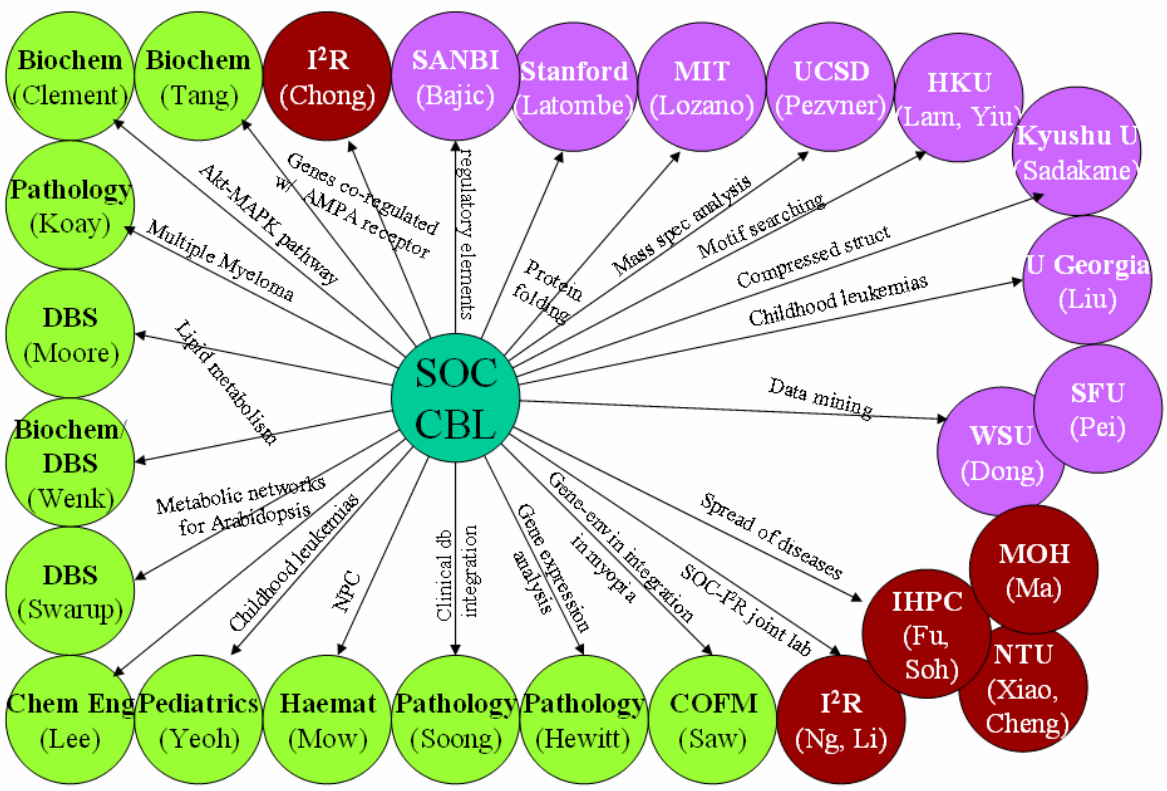
Achievements

We have developed a decompositional approach to parameter estimation for bio-dynamical systems based on hybrid functional Petri net. It exploits the structure of a large pathway model to break it into smaller components, whose parameters can then be estimated independently. This leads to significant improvements in computational efficiency. We have tested our approach on a detailed model of the Akt and MAPK pathways with two known and one hypothesized crosstalk mechanisms. Our simulation results exhibit good correlation with experimental data, and they yield positive evidence in support of the hypothesized crosstalk between the two pathways.

Relevant Publications

1. G. Koh, H. F. C. Teong, M.-V. Clement, D. Hsu, P. S. Thiagarajan. **A Decompositional Approach to Parameter Estimation in Pathway Modeling: A Case Study of the Akt and MAPK Pathways and their Crosstalk**. *Bioinformatics*, 22:e271--e280, 2006.

List of Collaborations



List of Faculty Members in Computational Biology

1. David HSU
2. Wynne HSU
3. LEE Mong LI
4. LEONG Hon Wai
5. OOI Beng Chin
6. SUNG Wing-Kin
7. TAN Kian Lee
8. P. S. Thiagarajan
9. Anthony TUNG
10. WONG Limsoon

Awards Received

- Sung Wing-Kin and his collaborators at the Genome Institute of Singapore were conferred the National Science Award in October 2006 for their innovative work in developing the Paired End diTag sequencing technology for comprehensive characterisation of the human genome and transcriptome.

- Wong Limsoon was conferred a Singapore Youth Award Medal of commendation in July 2006 for his sustained contributions to science and technology.


Journals and Proceedings Edited

Drug Discovery Today

Bioinformatics


JBCB

APBC'06

PRIB'06

Activities organized

- 5th Korea-Singapore Workshop on Bioinformatics and NLP, Feb 2006
- IMS Workshop on BioAlgorithmics, July 2006
- 3rd RECOMB Satellite Workshop on Regulatory Genomics, July 2006

Compiled by WONG Limsoon with inputs from CBL members
3 May 2007