Brief Overview of SOC's Computational Biology Lab

http://www.comp.nus.edu.sg/~cbl

10 May 2011

Computational Biology

- **Aims**
  - Improve understanding of molecular circuits
  - Deliver better diagnosis and treatment of diseases

- **Modeling & Simulation**


David Hsu


P.S. Thiagarajan

- **Combinatoric Algo**


Leong Hon Wai


Ken Sung

- **DB & Knowledge Discovery**


Wong Limsoon


Tan Kian Lee


Anthony Tung


Wynne Hsu


Lee Mong Li

Recent Honours

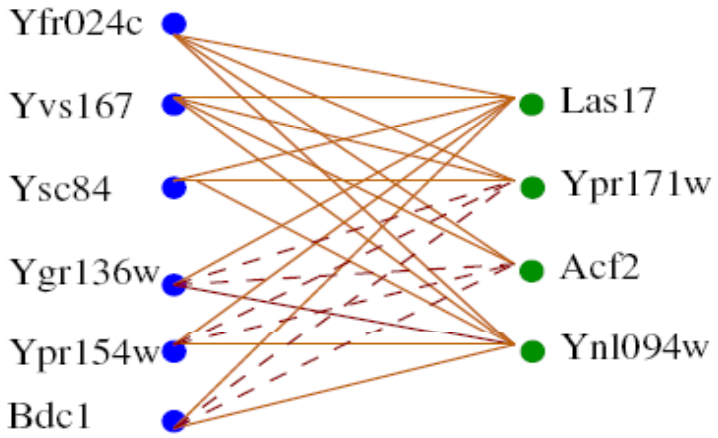
- **Ken Sung**
  - 2008 NUS Young Researcher Award: Contribution to research in algorithm & computational biology
  - 2006 Singapore National Science Award: Paired End diTag sequencing technology

- **Limsoon Wong**
  - 2006 Singapore Youth Award Medal of Commendation: Sustained contributions to science & technology
  - 2003 Far Eastern Economic Review Asian Innovation Gold Award: A simple test for childhood leukaemia

DREAM Challenge 2007

- **5 bioinformatics challenges**
- **Participants must predict the answer using bioinformatics methods**
- **We participated in 2 challenges and we were the best performers in both**

- **Challenge 1: BCL6 target genes finding**
  - Lee et al., *Ann NY Acad Sci*, 2009
- **Challenge 2: PPI subnetwork prediction**
  - Chua et al., *Ann NY Acad Sci*, 2009

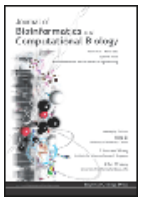


Recent Professional Activities

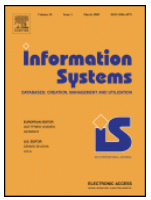
- **Journals edited:**

DDT

Bioinformatics



JBCB

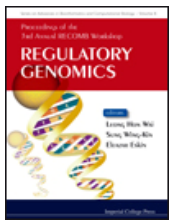


Info Systems

TCBB

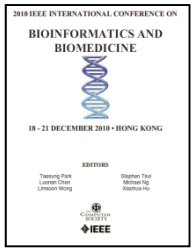
JOURNAL OF BIOMEDICAL SEMANTICS

- **Books/Proceedings edited:**

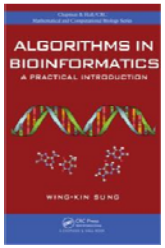


3rd Regulatory Genomics

RECOMB'08



BIBM2010



Algo in Bioinfo

- **Annually involve in ~10 bioinformatics conf prog & org committees**
  - RECOMB, ISMB, CSB, GIW, APBC, ACM BCB, IEEE BIBM, …
- **Annually publish ~30 papers**
  - Bioinformatics, JCB, BMC, JBCB, PLoS CB, EMBO, NAR, Cell, JBC, TKDE, VLDBJ, SIAM J Comput, Clin Cancer Res, JPR, …
- **Annually give ~10 keynotes & invited talks in conferences**

Recent Conferences Hosted

- **International Conferences**
  - 18th Intl Conf on Genome Informatics (GIW2007)
  - 2nd Intl Symp on Languages in Biology and Medicine (LBM2007)
  - 6th Assoc of Asian Societies for Bioinformatics Symp (AASBi2007)
  - 12th Intl Conf on Research in Computational Molecular Cell Biology (RECOMB2008)
- **Regional Workshops**
  - 1st Japan-Singapore Workshop on Computational Systems Biology (2008)
  - 8th Korea-Singapore Workshop on Bioinformatics & NLP (KSW2008)
  - 3rd Japan-Singapore Workshop on Computational Systems Biology (2011)
  - 2nd IPM-NUS Workshop on Computational Biology (2011)

Main Courses Taught

- **CS2220 Introduction to Computational Biology**
  - Understand bioinformatics problems; interpretational skills
- **CS3225 Combinatorial Methods in Bioinformatics**
- **CS4220 Knowledge Discovery Methods in Bioinformatics**
  - Clustering; classification; association rules; SVM; HMM; Mining of seq, trees, & graphs
- **CS5238 Advanced Combinatorial Methods in Bioinformatics**
  - Seq alignment, whole-genome alignment, suffix tree, seq indexing, motif finding, RNA sec struct prediction, phylogeny reconstruction
- **CS6280 Computational Systems Biology**
  - Dynamics of biochemical and signaling networks; modeling, simulating, & analyzing them
- **~15 students a year in NUS undergrad comp bio prog**

Recent Comp Bio PhD Students

http://www.comp.nus.edu.sg/~cbl/theses,html

- **2005: 3 PhD's awarded**
- **2006: 4 PhD's awarded**
- **2007: 4 PhD's awarded**
- **2008: 8 PhD's awarded**
- **2009: 4 PhD's awarded**
- **2010: 3 PhD's awarded**
- **2011: 3-5 PhD's expected**

- **Mengling Feng (2010)**
  - Frequent pattern space maintenance: Theories and algos
  - RF at A*STAR I^2R
- **Donny Soh (2010)**
  - Understanding pathways
  - RF at A*STAR I^2R
- **Charlie Lee (2010)**
  - Bioinformatics applications for virology research
  - RF at A*STAR GIS
- **Hugo Willy (2011)**
  - Interaction motif inference from biomolecular interactions
  - RF at NUS SOC
- **Bing Liu (2011)**
  - Probabilistic approx and analysis techniques for biopathway models
  - RF at NUS SOC
- **Brandon Ooi (2011)**
  - Molecular and computational approaches to understanding keloid scarring
  - Acad staff at Republic Poly

Research Directions

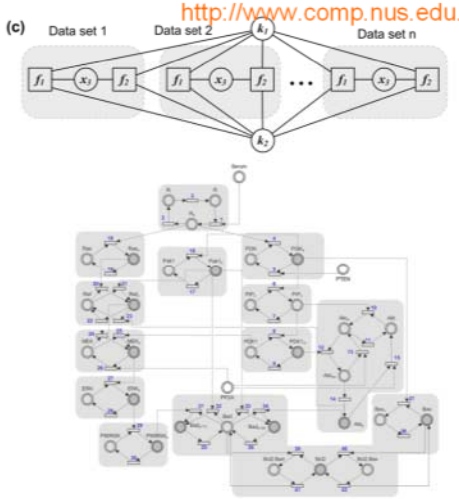
Recently Funded Projects

- **Supporting diagnostic data mining via exploratory hypothesis testing and analysis**, $655k, SERC PSF
- **Construction of reliable protein interactomes for infectious diseases**, $850k, MOE T2
- **Decomposition and Composition of Large Signalling Pathway Models with Emphasis on Parameter Estimation**, $487k, MOE AcRF T2
- **Extracting biological signals from second generation sequencing,** $888k, MOE T2
- **Lipidomics – Novel Tools and Applications,** $1m of $8m, NRF CRP
- **Total new funding**
  - $2.88m + $1m of 8m

Systems Biology Modeling and Simulation

Incremental Bio-pathway Modeling

http://www.comp.nus.edu.sg/~rpsysbio



- Model construction is an incremental process
- Factor graphs (prob graphical model) is used to represent pathway parameter estimates
- Temporal composition
  - Model refinement thru data integration
- Spatial composition
  - Model composition & expansion

Geoffrey Koh et al. RECOMB 2010

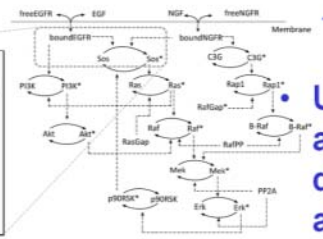
Approximation of ODE-Based Biological Pathway Dynamics

Probabilistic approximations of ODEs based bio-pathway dynamics*

Bing Liu[a,*], David Hsu[a,b], P.S. Thiagarajan[a,b]

ARTICLE INFO

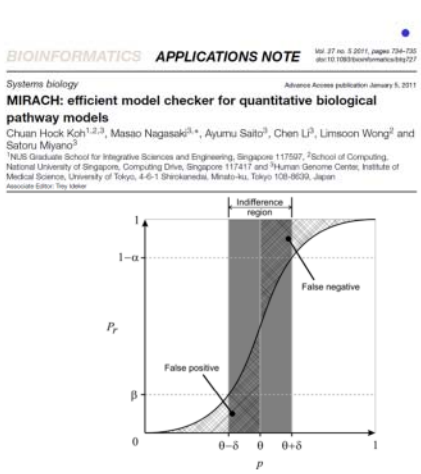
ABSTRACT



- Biochem networks are often modeled by ODEs
  - Simulations needed to perform analyses
  - # of simulations can become very large.
- Use discrete prob approx of ODEs dynamics to get around the problem

Liu et al. Theoret Comp Sci. 412:2188-2206, 2011

Model Checking Biological Pathways



BIOINFORMATICS APPLICATIONS NOTE

Systems biology

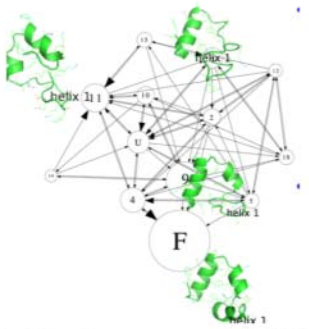
MIRACH: efficient model checker for quantitative biological pathway models

Chuan Hock Koh[1,2,3], Masao Nagasaki[3,*], Ayumu Saito[3], Chen Li[3], Limsoon Wong[2] and Satoru Miyano[3]

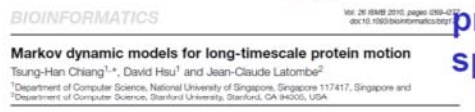
- MIRACH, a statistical online model checker for biopathway models
  - Support PLTL formalisms for expressing properties to be checked
  - Integrated w/ HFPNe simulation engine for fast on-the-fly model checking
  - Support pathway models written in CSML or SBML

Koh et al, Bioinformatics, 27:734-735, 2011

Markov Dynamic Model for Long-Timescale Protein Motion



- A probabilistic model of protein dynamics
  - Compact
  - Explicit
  - Explanatory
- A state is a probability distribution over protein conformation space


BIOINFORMATICS

Markov dynamic models for long-timescale protein motion

Tsung-Han Chiang[1,*], David Hsu[1] and Jean-Claude Latombe[2]

[1]Department of Computer Science, National University of Singapore, Singapore 117417, Singapore and [2]Department of Computer Science, Stanford University, Stanford, CA 94305, USA

Chiang et al, Bioinformatics, 26:i269-i277, 2010

Protein Structure, Folding, & Motion

Protein Flexible Region Identification

- Conformational changes play critical role in biological functions
- Can't compare backbone torsion angles due to noise in X-ray & NMR data
- Develop techniques to distinguish genuine conformational change from noise

⇒ Accurate identification of flexible vs rigid regions in proteins

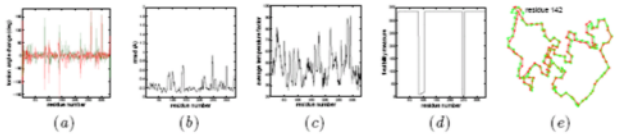
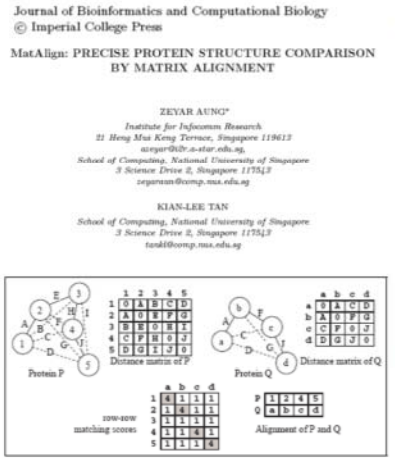


Fig. 1. Various methods for detecting flexibility in the N-lobe of lactoferrin. (*a*) Torsion angle differences. (*b*) The minimum RMSD for 5-residue fragments centered at each residue. (*c*) Average temperature factors from X-ray crystallography data. (*d*) Our new algorithm. For (*a*)–(*c*), large absolute values indicate flexible regions. For (*d*), small values indicate flexible regions. (*e*) Superimposition of the two conformations (in red and green, respectively) for the 40-residue fragment centered around residue 142.

Nagham et al, JCB, 15:813-828, 2008

Precise Structure Comparison


Journal of Bioinformatics and Computational Biology
© Imperial College Press

MatAlign: PRECISE PROTEIN STRUCTURE COMPARISON BY MATRIX ALIGNMENT

ZEYAR AUNG*
Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
azeyar@i2r.a-star.edu.sg,
School of Computing, National University of Singapore
3 Science Drive 2, Singapore 117543
zeyarzan@comp.nus.edu.sg

KIAN-LEE TAN
School of Computing, National University of Singapore
3 Science Drive 2, Singapore 117543
tankl@comp.nus.edu.sg

- MatAlign
  - Detailed struct alignment thru alignment of 2D dist matrix & iterative refinements
  - Provide better alignment scores than DALI & CE in majority of cases
  - 4 times faster than DALI, and has about the same speed as CE

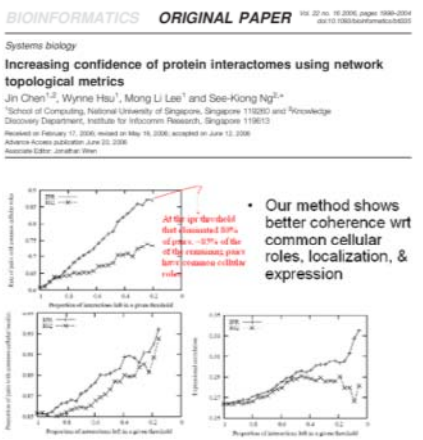
⇒ Significantly speed up searching of protein sequences and structures w/o sacrificing accuracy

Aung et al, JBCB, 4:1197-1216, 2006

Protein Interactions Reliability

- Protein-protein interaction expts have ~50% errors
- True interactions seem to exhibit certain topologies and motifs that can be modeled
- Develop computational methods to detect false positives
- Develop computational methods to detect false negatives

⇒ Robust and powerful system to identify protein-protein interactions in noisy expts


BIOINFORMATICS ORIGINAL PAPER

Systems biology

Increasing confidence of protein interactomes using network topological metrics

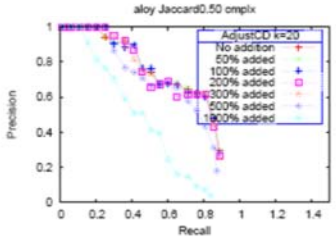
Jin Chen[1,2], Wynne Hsu[1], Mong Li Lee[1] and See-Kiong Ng[2,*]

- Our method shows better coherence wrt common cellular roles, localization, & expression

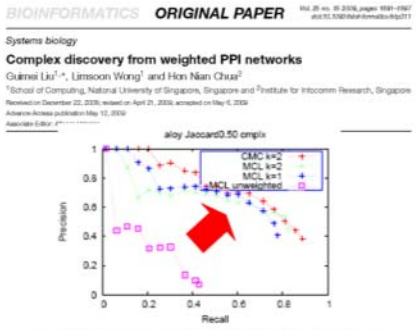
Chen et al, *Bioinformatics*, 22:1998-2004, 2006

Protein Complex Prediction

- Reliable cleansing of PPI network by expectation maximization of score based on shared interaction partners



- Robust up to 500% noise PPIs


BIOINFORMATICS ORIGINAL PAPER

Systems biology

Complex discovery from weighted PPI networks

Guimei Liu[1,*], Limsoon Wong[1] and Hon Nian Chua[2]

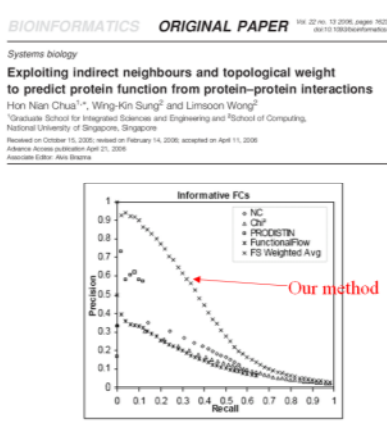
⇒ Uniformly improved existing protein complex prediction methods (MCL)

⇒ New robust system for protein complex prediction (CMC)

Liu et al, *Bioinformatics*, 25:1891-1897, 2009

Protein Function & Protein Complex Prediction Using PPI Networks

Protein Function Prediction


BIOINFORMATICS ORIGINAL PAPER

Systems biology

Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions

Hon Nian Chua[1,*], Wing-Kin Sung[2] and Limsoon Wong[2]

- How significant is functional association between level-2 neighbors?
- How can they be exploited for protein function prediction?
- How to integrate protein interaction info with other info to improve protein function prediction?

⇒ Robust and powerful system to predict protein functions based on PPIs

Chua et al, *Bioinformatics*, 22:1623-1630, 2006

Cloud Computing for GWAS

BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 8 2011, pages 1045–1051
doi:10.1093/bioinformatics/btr09

Genome analysis Advance Access publication March 2, 201

eCEO: an efficient Cloud Epistasis cOmputing model in genome-wide association study

Zhengkui Wang[1,*], Yue Wang[1], Kian-Lee Tan[1,2], Limsoon Wong[1,2] and Divyakant Agrawal[3]

[1]NUS Graduate School for Integrative Sciences and Engineering, [2]Department of Computer Science, School of Computing, National University of Singapore, Singapore and [3]Department of Computer Science, University of California, Santa Barbara, 93106-5110, USA

Associate Editor: Martin Bishop

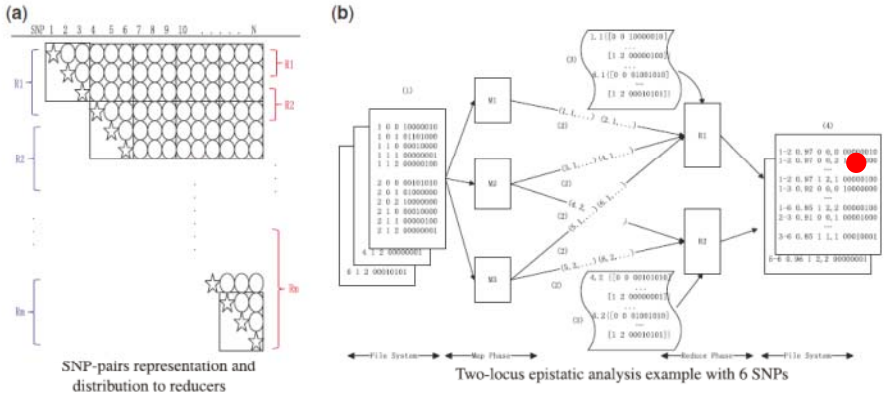


Fig. 3. Parallel distribution models and example of two-locus epistatic analysis using *eCEO* model.

- **Cloud-based epistasis computing model for large-scale GWAS**
  - Efficient
  - Flexible
  - Scalable
  - Practical

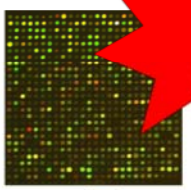
eCEO can do pairwise testing of 500k SNPs in <9 hrs using 40 nodes


Ken Sung

Better procedure for sequencing H1N1 in human samples

Human sample with H1N1 virus

This chip was designed by Ken Sung & his team



Almost all Singapore H1N1 sequences were sequenced using this chip! The result was used by MOH for decision making.

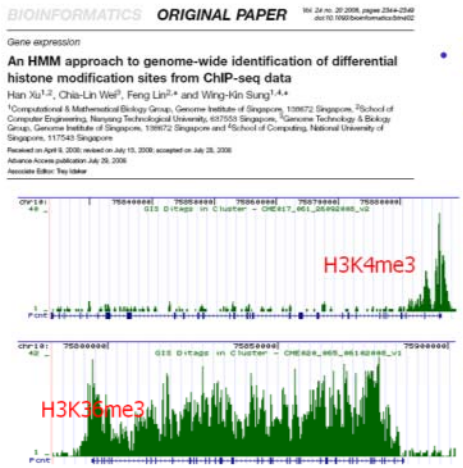
Lee et al, Nucleic Acids Research, 2010

>H1N1 sequence

……ACGTCAGGTCATGCATGGTCAAG……

Next-Gen Sequencing & Transcriptomics

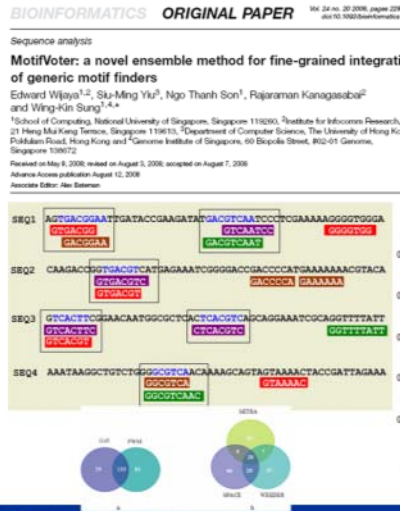
Genome-Wide Identification of Differential Histone Modification Sites from ChIP-Seq Data



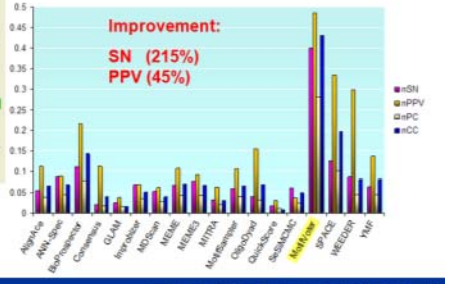
- First method to identify broad histone modifications in genome-wide scale from ChIP-seq data
- Based on Hidden Markov Model (HMM)
- The method also suggested that gene expression can be predicted by K4 and K36

Xu et al, *Bioinformatics*, 24:2344-2349, 2008

Ensemble Method for Motif Finding



- Many motif finders exist
- Different motif finders give different results
- Idea: Ensemble output of different motif finders

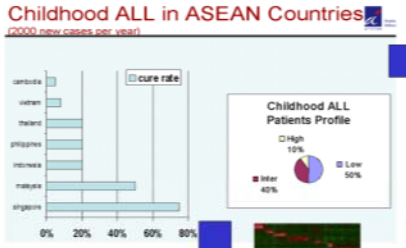


Wijaya et al, *Bioinformatics*, 24:2288-2295, 2008

Network-Based Gene Expression & Proteomic Profile Analysis

Wong Limsoon

Better disease diagnosis & treatment



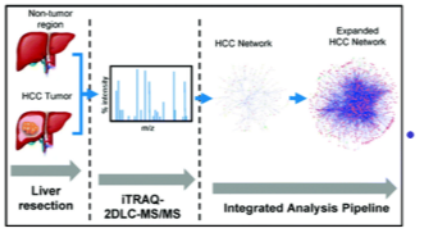
Conventional Tx:
- **intermediate intensity to all**
- **⇒ 10% suffers relapse**
- **⇒ 50% suffers side effects**
- **⇒ costs US$150m/yr**

Bioinformatics-optimized Tx:
- **high intensity to 10%**
- **intermediate intensity to 40%**
- **low intensity to 50%**
- **costs US$100m/yr**

- **High cure rate of 80%**
- **Less relapse**
- **Less side effects**
- **Save US$51.6m/yr**

Yeoh et al, *CANCER CELL*, 2002

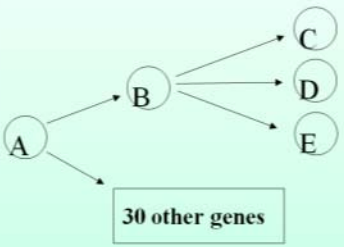
Network-Based Proteomic Profile Analysis



- **Current high thru'put MS led to noisy proteomic profiles**
- **Proteomic expansion pipeline (PEP)**
  - Expansion by 1st-deg PPI partners improves coverage greatly
- **Proteomic signature profiling (PSP)**
  - Threshold-free approach to cancer proteomics

Goh et al, J Proteome Res, 10:2261-2272, 2011

Subnet-Based Gene Expr Analysis



A branch within pathway consisting of genes A, B, C, D and E are high in phenotype *D*

Genes C, D and E not high in phenotype *~D*

30 other genes not diff expressed

Conventional techniques: Entire network is likely to be missed

- **SNet: Capture subnetwork branch within pathway**
  - Highly reproducible large subnets differentially expressed betw patient phenotypes

Fast DNA Alignment

BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 6 2008, pages 791–797
doi:10.1093/bioinformatics/btn032

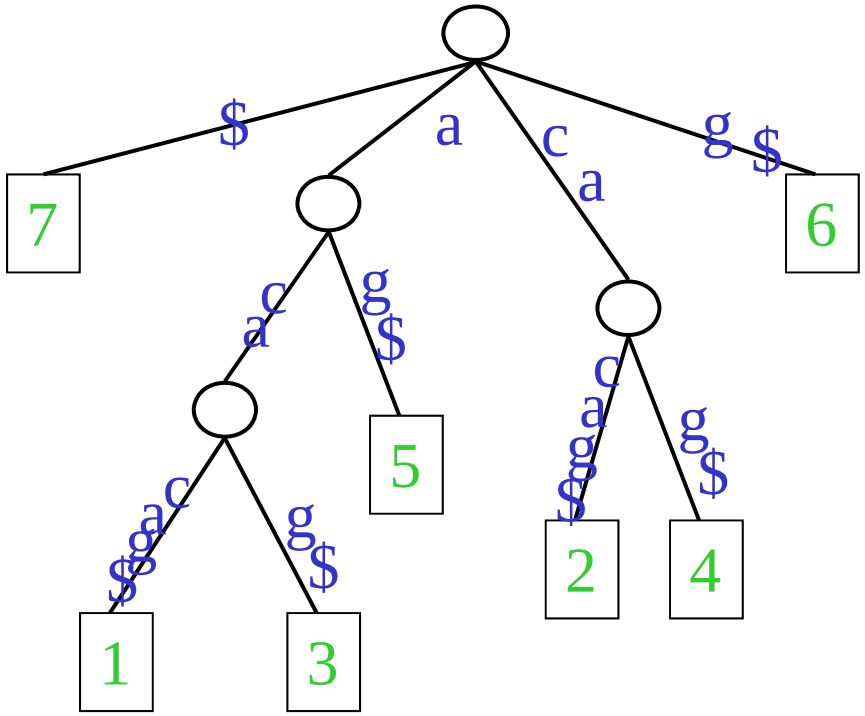
Sequence analysis

Compressed indexing and local alignment of DNA

T. W. Lam[1,*], W. K. Sung[2], S. L. Tam[1], C. K. Wong[1] and S. M. Yiu[1]

[1]Department of Computer Science, University of Hong Kong, Hong Kong, China and [2]Department of Computer Science, National University of Singapore, Singapore

Received on August 29, 2007; revised on December 8, 2007; accepted on January 22, 2008
Advance Access publication January 28, 2008
Associate Editor: Thomas Lengauer



- **BLAST is one of the best methods for identify approx matching in a large seq db**
- **However, it is a heuristics. It will miss answers**
- **We introduce meaningful alignment based on compressed suffix tree**

⇒ **New DNA alignment method that does not miss answers and is as fast as BLAST**

Query length	100	200	500	1 K	2 K
BWT-SW average time (s)	1.91	4.02	9.89	18.86	35.93
Smith–Waterman average time (K)	5.1	10.0	23.9	45.1	97.8
BLAST average time	9.7	12.58	12.52	15.23	15.82

Lam et al, *Bioinformatics*, 24:791-797, 2008

Any Question?

Contact:

Professor Wong Limsoon

http://www.comp.nus.edu.sg/~wongls