# Brief Overview of SOC's Computational Biology Lab

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

## Lab Coordinator: LimSoon *WONG*

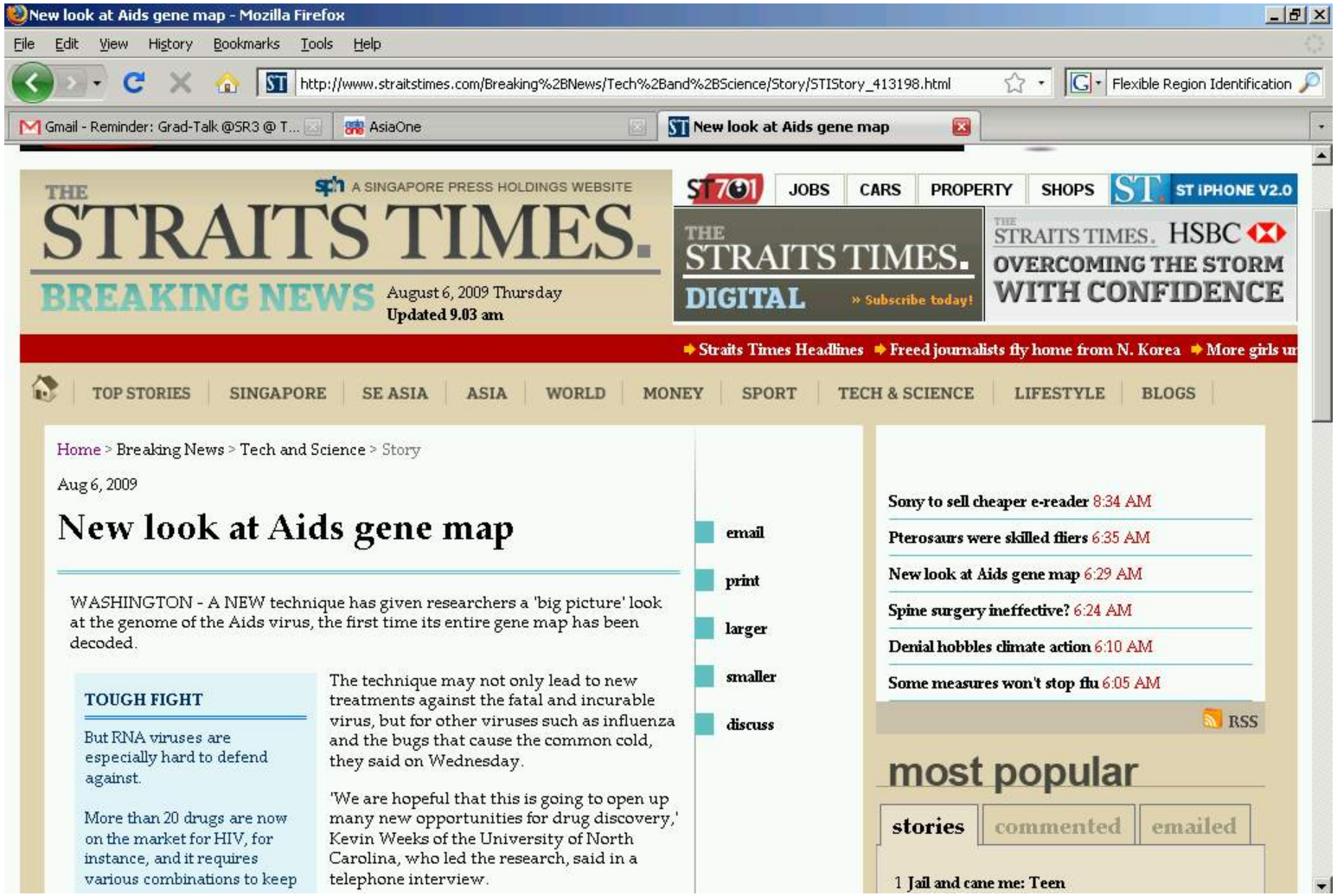
**Presentation by: Hon Wai *LEONG***

***School of Computing***

***National University of Singapore***

**06 August 2009**

# Hot from this morning's news…




Copyright © 2009 by Limsoon Wong

Aug 6, 2009

# New look at Aids gene map

WASHINGTON – A NEW technique has given researchers a 'big picture' look at the genome of the Aids virus, the first time its entire gene map has been decoded.

**TOUGH FIGHT**

But RNA viruses are especially hard to defend against.

More than 20 drugs are now on the market for HIV, for instance, and it requires various combinations to keep it in check.

... more

The technique may not only lead to new treatments against the fatal and incurable virus, but for other viruses such as influenza and the bugs that cause the common cold, they said on Wednesday.

'We are hopeful that this is going to open up many new opportunities for drug discovery,' Kevin Weeks of the University of North Carolina, who led the research, said in a telephone interview.

The human immunodeficiency virus or HIV is what is known as an RNA virus. Like influenza, polio and many viruses that cause colds, it uses RNA instead of DNA as its map when carrying out functions.

DNA depends on building blocks called nucleotides to carry information on its two strands. These are the familiar A, C, T and G of the genetic code. RNA has just one strand and depends on complex folding patterns to carry information, as well as nucleotides.

'There is so much structure in the HIV RNA genome that it almost certainly plays a previously unappreciated role in the expression of the genetic code,' Dr Weeks said.

His team developed a new chemical method called SHAPE to make an image not only of the RNA's nucleotides, but of the shapes and folds of the RNA strands.

Other imaging methods such as X-ray crystallography can capture the precise position of each atom, but only one small area at a time. SHAPE gets a bigger picture, but not at the atomic level, Dr Weeks said.

'The technique is thus akin to zooming out on a map and getting a broader view of the landscape at the expense of fine details,' Hashim Al-Hashimi of the University of Michigan wrote in a commentary on the findings, published in Nature.

This, in turn, will help researchers make better drugs to treat such viruses, said Dr Weeks. New drugs are often engineered to fit into specific structures on a virus, blocking it from attaching to a cell, for instance, or gumming up its works so it cannot replicate. -- REUTERS

Copyright © 2009 by Limsoon Wong

# Research

**Aims**

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

**Research**

- **Combinatorics & Algorithms**
- **Database Technologies**
- **Knowledge Discovery Technologies**
- **Modeling, Simulation & Analysis**

**Applications**

- **Analysis of Seq Data**
- **Speciality Databases**
- **High-Throughput Expts**
- **Analysis of Clinical Data**
- **Analysis of Protein Structure & interactions**
- **Molecular Evolution**
- **Signaling pathways dynamics**

Copyright © 2009 by Limsoon Wong

# People


David Hsu


Ken Sung


Anthony Tung


Mong Li Lee


Wynne Hsu


Kian Lee Tan


Hon Wai Leong


Beng Chin Ooi


P.S. Thiagarajan


Limsoon Wong
(***Coordinator***)

- **Postdocs: 2**
- **Students: 31**
- **Alumni: 35**

Copyright © 2009 by Limsoon Wong

# 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

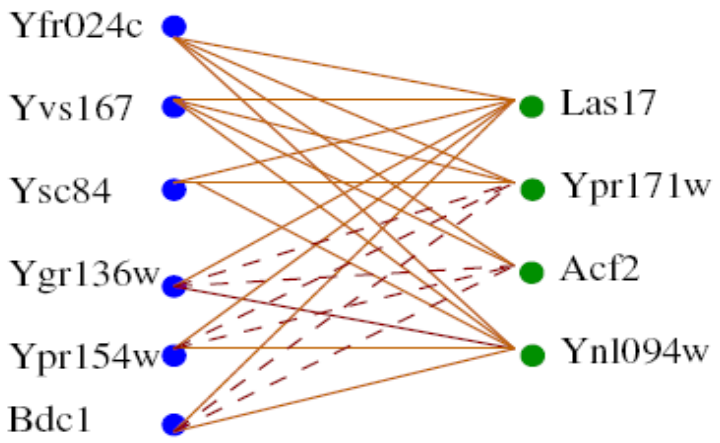
- **Limsoon Wong**
  - 2006 Singapore Youth Award Medal of Commendation: Sustained contributions to science & technology
  - 2003 Far Eastern Economic Review Asian Innovation Gold Award: A simple test for childhood leukaemia

Copyright © 2009 by Limsoon Wong

# DREAM Challenge 2007

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

- **Challenge 1: BCL6 target genes finding**
  - Charlie Lee et al.
- **Challenge 2: PPI subnetwork prediction**
  - Kenny Chua et al.

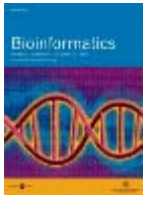



Copyright © 2009 by Limsoon Wong

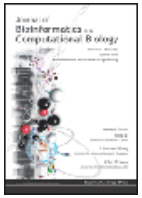
# Professional Activities in 2007/8

- **Journals edited:**


**DDT**



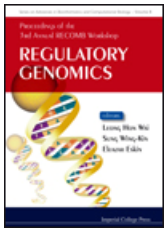

**Bioinformatics**



**JBCB**

BMC Research Notes

- **Books/Proceedings edited:**




**3rd Regulatory Genomics**


**GIW'07**


**RECOMB'08**

- **Involved in ~20 bioinformatics conf prog & org committees**
  - RECOMB07/08, ECCB07, ISMB07/08, CSB07/08, GIW07/08, APBC07/08, …
- **Published ~80 papers**
  - Bioinformatics, JCB, BMC, JBCB, TCBB, DDT, AJHG, Nature, Mol Cell, Genome Biology, Genome Res, Cell Stem Cell, …
- **~30 keynotes & invited talks in conferences**

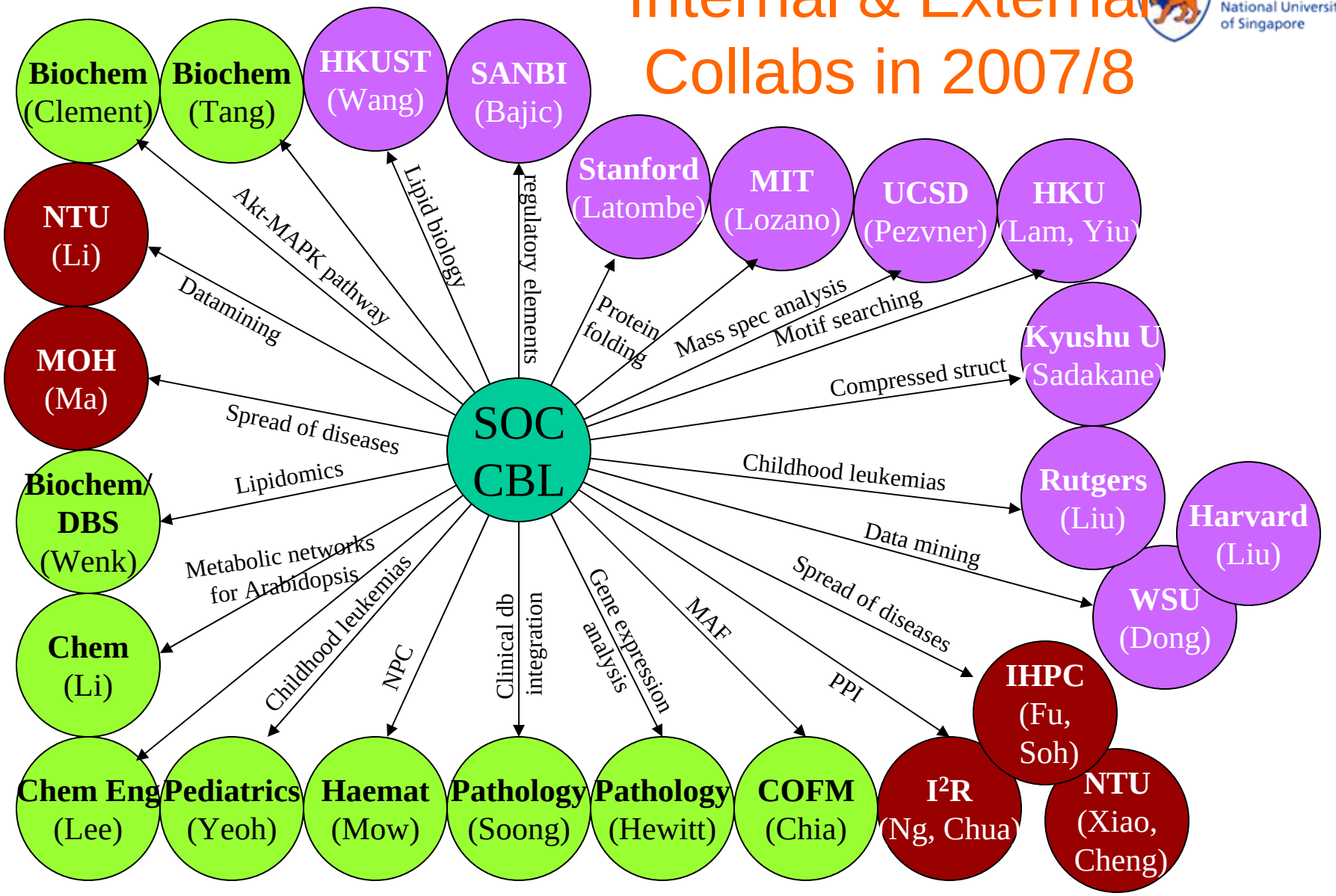
Copyright © 2009 by Limsoon Wong

# Conferences Hosted in 2007/8

- **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)**
- **1st Japan-Singapore Workshop on Computational Systems Biology (2008)**
- **8th Korea-Singapore Workshop on Bioinformatics & NLP (KSW2008)**

Copyright © 2009 by Limsoon Wong

# Internal & External Collabs in 2007/8




Copyright © 2009 by Limsoon Wong

# Main Courses Developed

- **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
- **Etc …**

Copyright © 2009 by Limsoon Wong

# Placement of Students in 2008

- **2005: 3 PhD's awarded**
- **2006: 4 PhD's awarded**
- **2007: 4 PhD's awarded**
- **2008: 8 PhD's awarded**

- **Kang Ning**
  - Algo for peptide and PTM …
  - PDF at Univ Michigan
- **V. S. Sundararajan**
  - Progressive data mining: …
  - RF at SANBI
- **Edward Wijaya**
  - Integrative methods for discovering…
  - PDF at JAIST
- **Hon Nian Chua**
  - Graph-based methods for protein function prediction
  - RF at A*STAR $I^2R$
- **Geoffrey Koh**
  - Pathway models decomposition …
  - RF at A*STAR BTI
- **Li Lin**
  - Efficient mining of haplotype …
  - Lecturer at SIM Univ
- **Stanley Ng**
  - Computational identification of novel microRNAs …
  - RF at A*STAR SIgN
- **Swee Seong Wong**
  - String matching …
  - Sr Assoc Scientist at LSCDD

Copyright © 2009 by Limsoon Wong

# Research Highlight

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

BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 20 2008, pages 2344–2349
doi:10.1093/bioinformatics/btn402

*Gene expression*

**An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data**

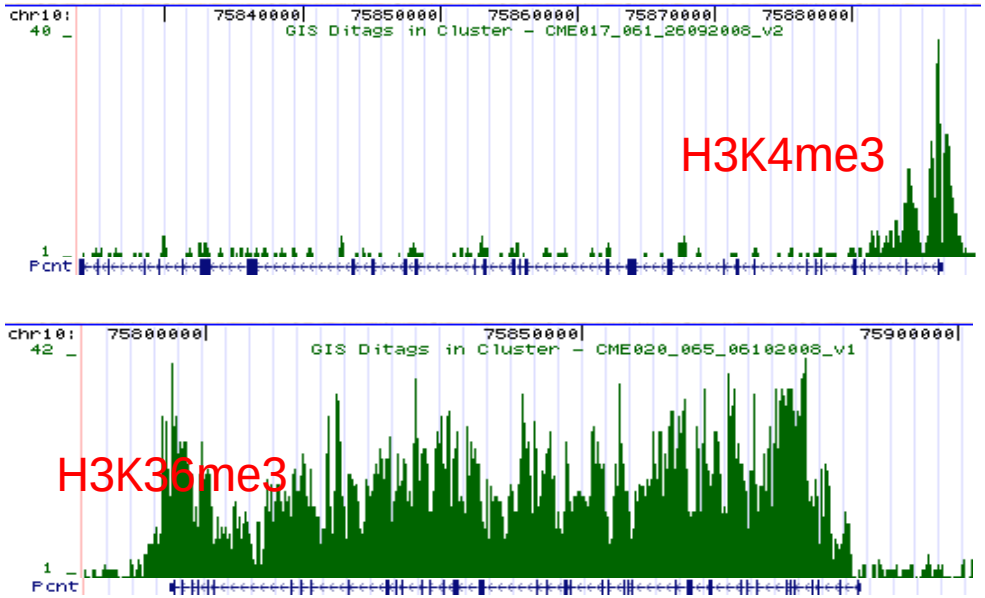
Han Xu[1,2], Chia-Lin Wei[3], Feng Lin[2,*] and Wing-Kin Sung[1,4,*]

[1]Computational & Mathematical Biology Group, Genome Institute of Singapore, 138672 Singapore, [2]School of Computer Engineering, Nanyang Technological University, 637553 Singapore, [3]Genome Technology & Biology Group, Genome Institute of Singapore, 138672 Singapore and [4]School of Computing, National University of Singapore, 117543 Singapore

Received on April 9, 2008; revised on July 13, 2008; accepted on July 28, 2008

Advance Access publication July 29, 2008

Associate Editor: Trey Ideker




- **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**

Copyright © 2009 by Limsoon Wong

# Ensemble Method for Motif Finding

BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 20 2008, pages 2288–2295
doi:10.1093/bioinformatics/btn420

*Sequence analysis*

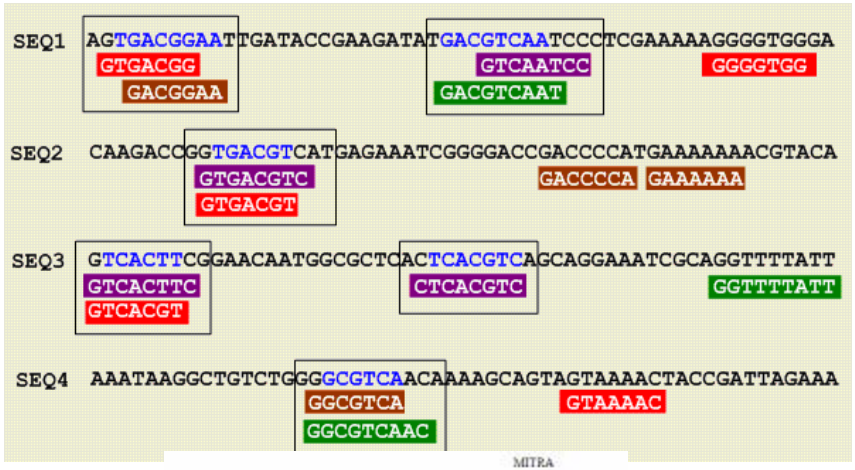
**MotifVoter: a novel ensemble method for fine-grained integration of generic motif finders**

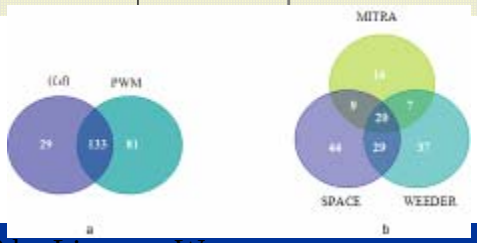
Edward Wijaya[1,2], Siu-Ming Yiu[3], Ngo Thanh Son[1], Rajaraman Kanagasabai[2] and Wing-Kin Sung[1,4,*]

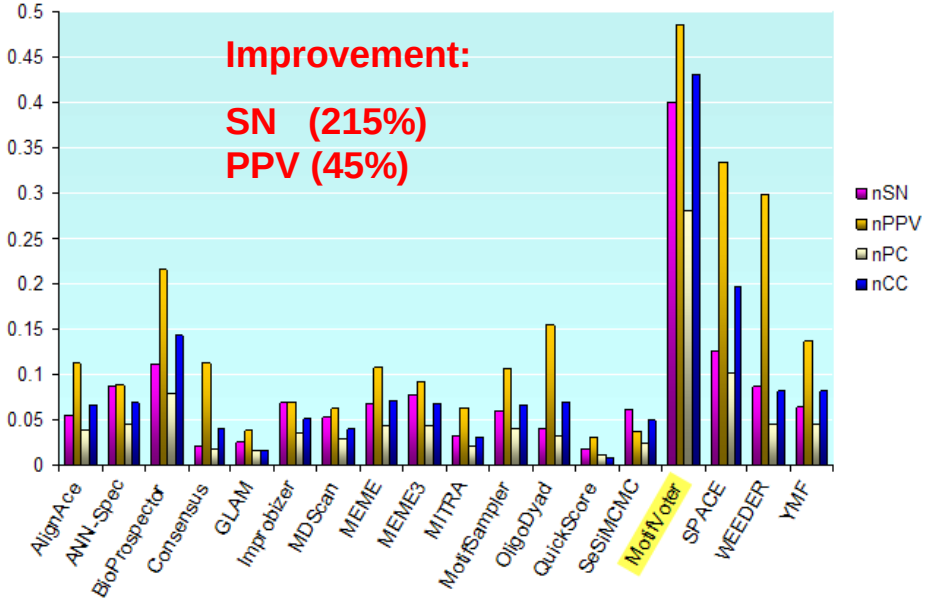
[1]School of Computing, National University of Singapore, Singapore 119260, [2]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, [3]Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong and [4]Genome Institute of Singapore, 60 Biopolis Street, #02-01 Genome, Singapore 138672

Received on May 9, 2008; revised on August 3, 2008; accepted on August 7, 2008
Advance Access publication August 12, 2008
Associate Editor: Alex Bateman

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




Copyright © 2009 by Limsoon Wong

# Fast DNA Alignment

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

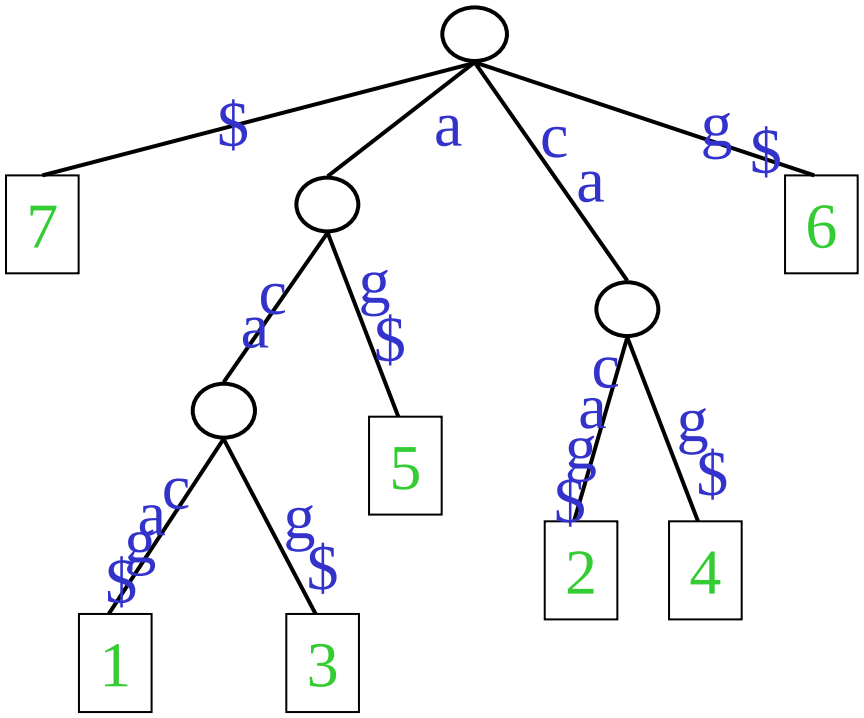
Sequence analysis

**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 |

Copyright © 2009 by Limsoon Wong

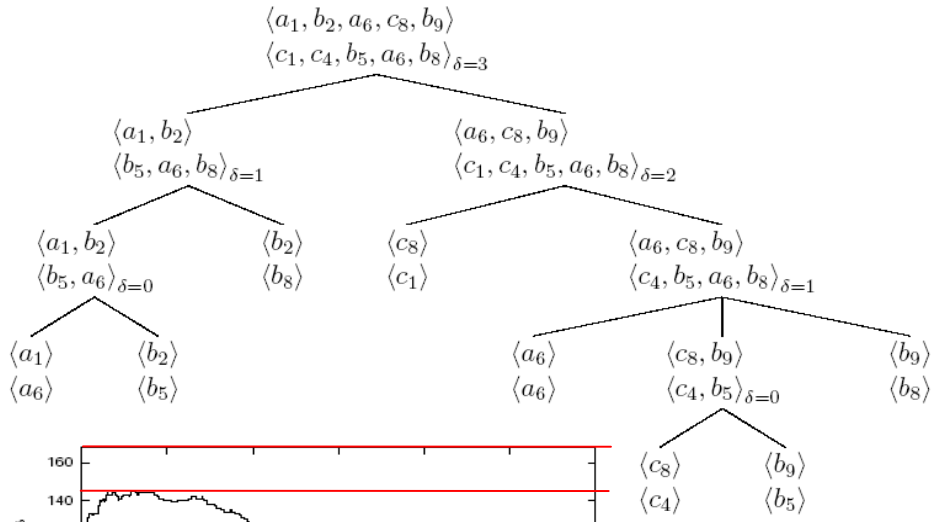
Lam et al, *Bioinformatics*, 24:791-797, 2008

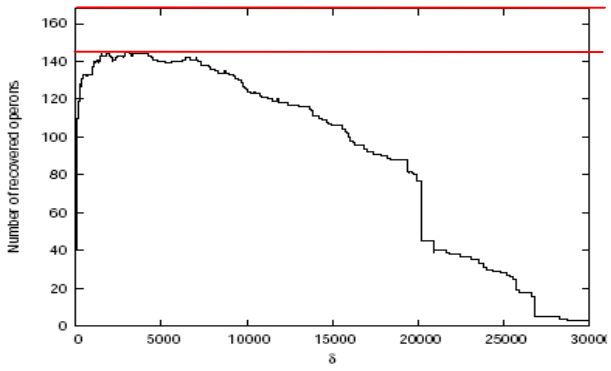
# Conserved Gene Clusters Identification Using Gene Team Tree


C.E. Nelson and S. Vialette (Eds.): RECOMB-CG 2008, LNBI 5267, pp. 100–112, 2008.
c Springer-Verlag Berlin Heidelberg 2008

Gene Team Tree:
A Compact Representation of All Gene Teams

Melvin Zhang and Hon Wai Leong
School of Computing, National University of Singapore




- **How to find biologically significant conserved genes clusters?**
- **Current methods require specification of model parameters. Non-trivial in practice.**
- **Our method finds a hierarchical clustering tree over ALL parameter values.**

⇒ **More comprehensive coverage of significant gene clusters (15% more operons identified when comparing E.coli and B.subtilis)**

Copyright © 2009 by Limsoon Wong

# Protein Function Prediction from PPI

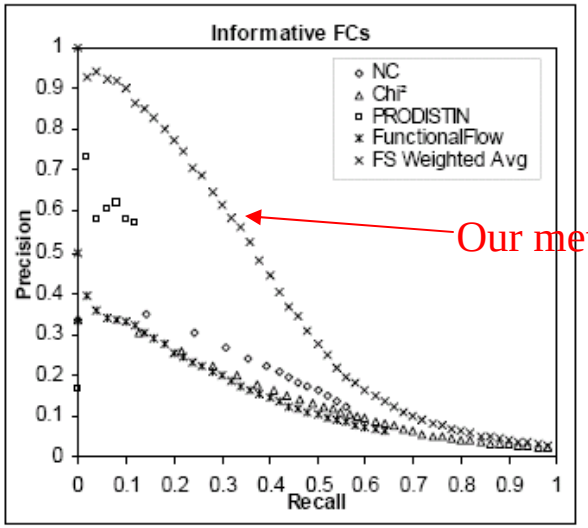
BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 13 2006, pages 1623–1630
doi:10.1093/bioinformatics/btl145

*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]
[1]Graduate School for Integrated Sciences and Engineering and [2]School of Computing, National University of Singapore, Singapore

Received on October 15, 2005; revised on February 14, 2006; accepted on April 11, 2006
Advance Access publication April 21, 2006
Associate Editor: Alvis Brazma




- **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**

Copyright © 2009 by Limsoon Wong

# Protein Function Prediction

BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 24 2007, pages 3364–3373
doi:10.1093/bioinformatics/btm520

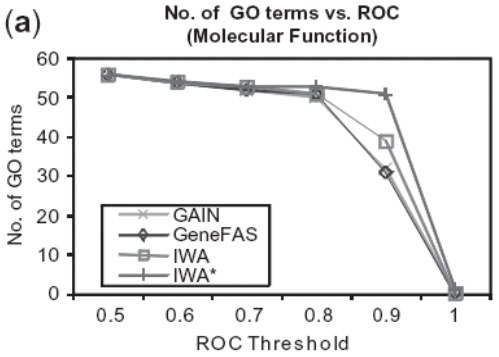
*Systems biology*

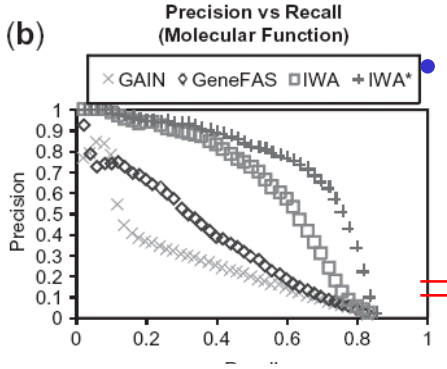
**An efficient strategy for extensive integration of diverse biological data for protein function prediction**

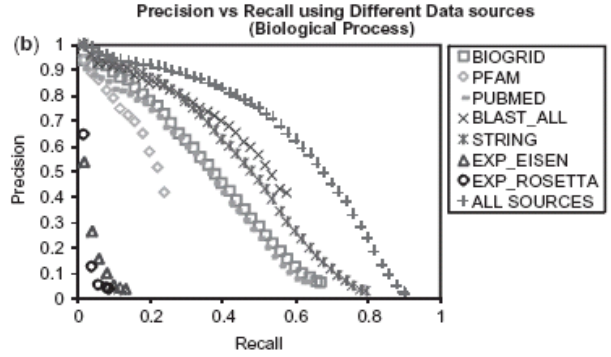
Hon Nian Chua[1,*], Wing-Kin Sung[2] and Limsoon Wong[2]
[1]Graduate School for Integrative Sciences and Engineering and [2]School of Computing, National University of Singapore, Singapore

Received on May 1, 2007; revised and accepted on October 12, 2007

Associate Editor: Chris Stoeckert



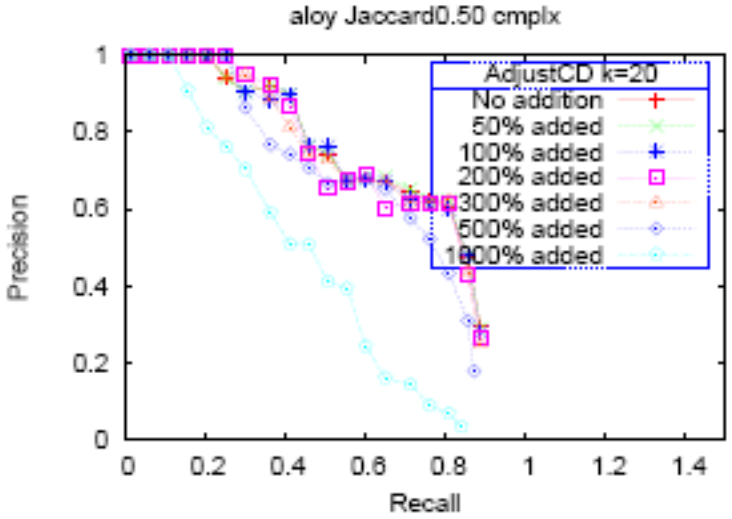

- **Simple effective framework for integrating large amt of diverse info for protein function prediction**
- **Exceptional performance compared to state of art**

⇒ **New robust system (IWA) to predict protein functions, even w/o sequence homology**

Copyright © 2009 by Limsoon Wong

# Protein Complex Prediction

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

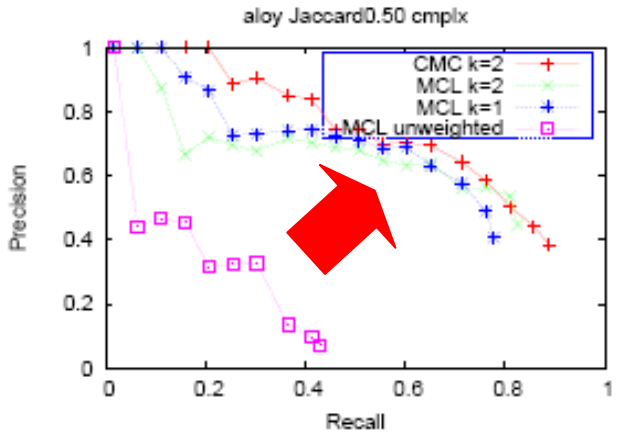



- **Robust up to 500% noise PPIs**

BIOINFORMATICS
Vol. 00 no. 00 2008
Pages 1–7

**Complex Discovery from Weighted PPI Networks**

Guimei Liu [1]*, Limsoon Wong [1], Hon Nian Chua [2]
[1]School of Computing, National University of Singapore and
[2]Institute for Infocomm Research, Singapore




⇒ **Uniformly improved existing protein complex prediction methods (MCL)**

⇒ **New robust system for protein complex prediction (CMC)**

Copyright © 2009 by Limsoon Wong

Liu et al, *Bioinformatics*, submiited, 2008

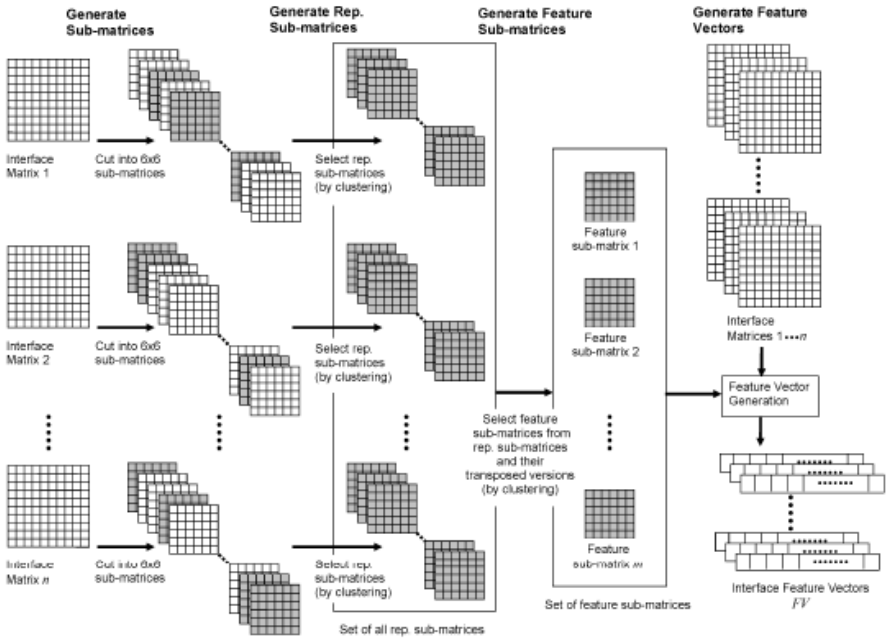
# Uncovering Structural Basis of PPI

Journal of Bioinformatics and Computational Biology
© Imperial College Press

PPiClust: EFFICIENT CLUSTERING OF 3-D PROTEIN–PROTEIN INTERACTION INTERFACES*

ZEYAR AUNG† SOON-HENG TAN‡ SEE-KIONG NG
*Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613*
{azeyar, shtan, skng}@i2r.a-star.edu.sg

KIAN-LEE TAN
*School of Computing, National University of Singapore, Law Link, Singapore 117590*
tankl@comp.nus.edu.sg




- **PPiClust**
  - Systematically encode, cluster, & analyse similar 3D interface patterns in protein complexes
  - Discover consistent and statistically significant clusters of interfaces
  - 8 hours vs 4 years of processing time compared to I2I-SiteEngine

⇒ **Efficiently detect spatially conserved but sequentially discontinuous biological motifs**

Copyright © 2009 by Limsoon Wong

Aung et al, *JBCB*, 6:415-433, 2008

# Protein Flexible Region Identification

*David Hsu*

- **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**

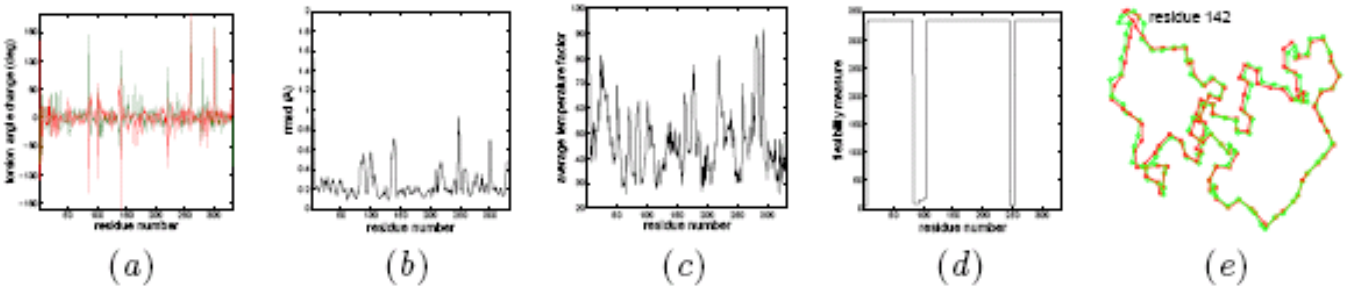



**Fig. 1.** Various methods for detecting flexibility in the N-lobe of lactoferrin. (*a*) Torsion angle differences. (*b*) The minimum RMSD for 5-residue fragments centered at each residue. (*c*) Average temperature factors from X-ray crystallography data. (*d*) Our new algorithm. For (*a*)–(*c*), large 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.

Copyright © 2009 by Limsoon Wong

Nigham et al, *JCB*, 15:813-828, 2008

# Peptide Sequencing using Multi-Charge MS/MS Spectra

NUS National University of Singapore

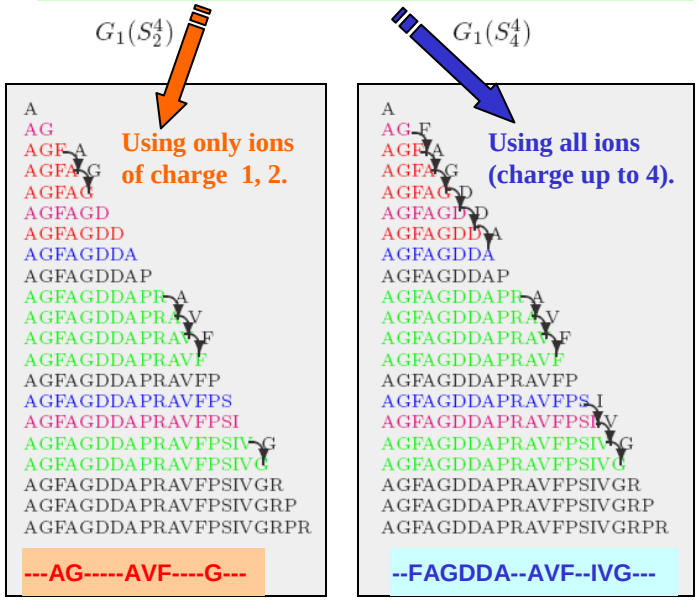
Journal of Bioinformatics and Computational Biology
Vol. 4, No. 6 (2006) 1329–1352
© Imperial College Press
Imperial College Press
www.icpress.co.uk

MODELING AND CHARACTERIZATION OF MULTI-CHARGE MASS SPECTRA FOR PEPTIDE SEQUENCING

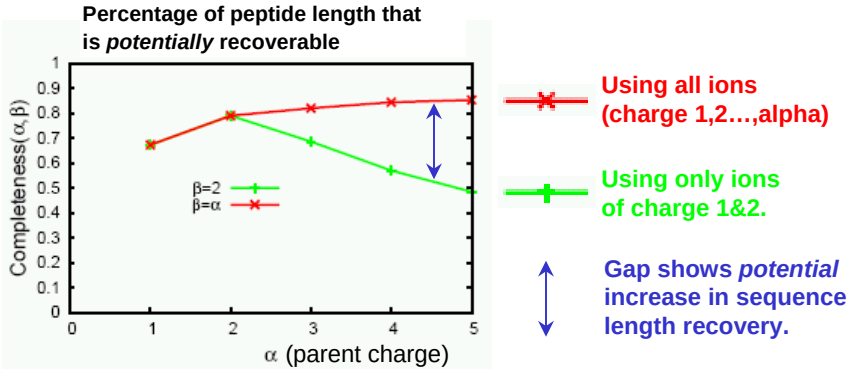
KET FAH CHONG*, KANG NING† and HON WAI LEONG†
*Department of Computer Science*
*National University of Singapore*

PAVEL PEVZNER
*Department of Computer Science & Engineering,*
*University of California, San Diego, La Jolla, CA 92093-0114*

**For the same MS/MS Spectrum with parent charge 4.**
**Original peptide: AGFAGDDAPRAVFPSIVGRPR**




- **Many sequencing methods considers only ions of charge 1 & 2.**
- **Can we recover more of the sequence with ions of higher charge (> 2)?**
- **Big *potential* gain in peptide recovery**

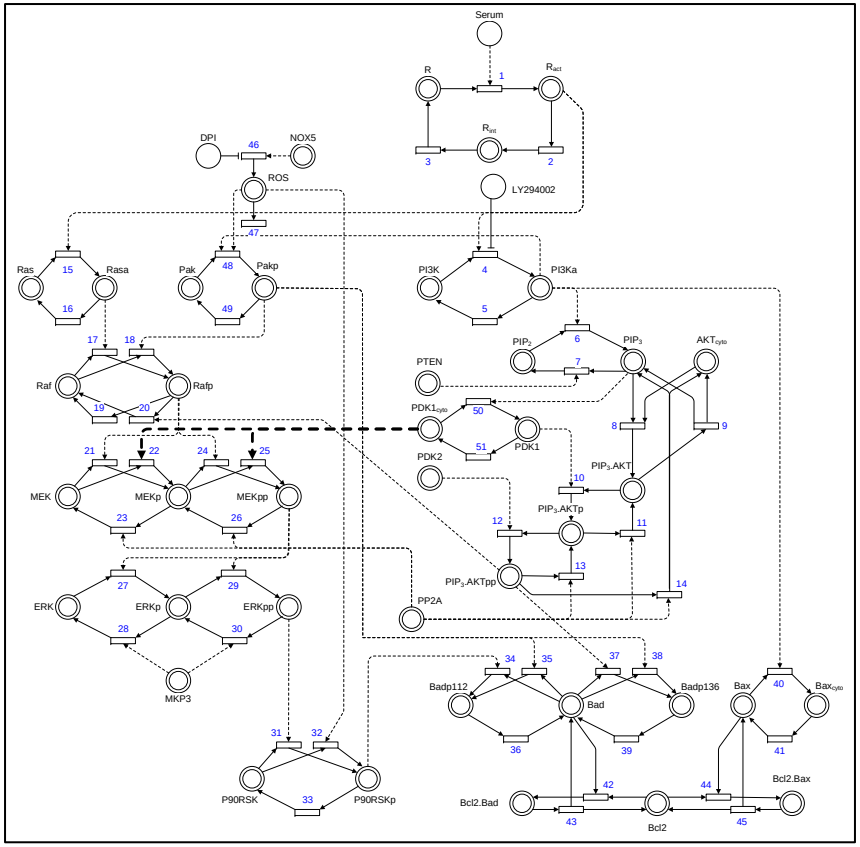



- **Can this improvement be realized?**
- **Even *simple* algorithm *using all ions* can improve peptide sequence recovery compared to existing methods.**

Copyright © 2009 by Limsoon Wong

# Parameter Estimation via Decomposition

- **Many bio-chemical reactions have unknown rate parameters; need to be estimated**
- **Decompose large pathway model into smaller "executable" models**
- **Estimate parameters for component models**
- **Compose component models using belief propagation (to reconcile conflicting parameter values of common portions)**




Copyright © 2009 by Limsoon Wong

Koh et al, *Bioinformatics*, 22:e271-e280, 2006

# Any Question?

**Contact:**

**Hon Wai *LEONG***
**Limsoon *WONG***
**{leonghw, wongls}@comp.nus.edu.sg**

# Applied Algorithms Research

- **Leong Hon Wai**
  - **Office: COM1 03-41**
  - **http://www.comp.nus.edu.sg/~leonghw/**
- **Research Lab:**
  - **Algorithms Lab (COM1 01-09)**
- **Applied Algorithms:**
  - **Design and Analysis of Algorithms**
  - **Algorithms for Transportation, Logistics and OR**
  - **Algorithms for Computational Biology**