SIX-MONTHLY PROGRESS REPORT

FOR I^{2}R-SOC JOINT LAB

Section A: To be completed by Principal Investigator

Reporting Period	1 February 2005 – 31 October 2005
Project Vote No. at SOC	R-252-000-172-593
Title of Project	Knowledge Discovery from Biological & Clinical Data
PIs	A/P Wynne Hsu, SOC Prof. Wong Limsoon, SOC/I^{2}R
Co-PIs	Dr. Lee Mong Li, SOC Dr. Ken Sung, SOC Prof. Vladimir Bajic, I^{2}R Prof. Vladimir Brusic, I^{2}R (Left for Univ of Queensland, Aug 05) A/Prof. Ng See Kiong, I^{2}R A/Prof. Li Jinyan, I^{2}R
Project Duration (Start Date – End Date)	1 July 2003 to 30 June 2006

1. Expenditure Level – Utilisation Rate

Vote	Original Grant	Revised Grant (if applicable)	Actual expenditure to date (exclude commitments)	% Utilization
EOM	$822,000	NA	$273,522.96	52.2%
OOE	$45,000	NA	$22,994.05	80.1%
EQPT	$70,00	NA	$15,230.77	26.44%
Total	$937,000	NA	$311,747.78	51.0%

Comments (to include explanation for major variations/virements; use additional pages if necessary)

The utilization for EOM for research scholars is lower than anticipated. This is because we are highly selective in our recruitment in order to ensure a good balance of locals, ASEAN, PRC and Indian candidates. This is aggravated by the global slowdown of students applying for graduate studies.

SoC has been very generous in providing high-end PCs to the research scholars. We purchased two 1TBytes storage systems but the amount is not significant due to the large drop in the storage system pricing.

2. Manpower Development - Project Staffing Status

Manpower Category	Planned Full-time	Actual Full-time
PhD student	10	3
Total	10	3

Comments (to include explanation for major deviations from approved targets and problems encountered; use additional pages if necessary)

PhD students recruited & funded for this project:

Postdoc recruited & funded for this project:

1. **Yang Liang-Huai (till 31 October 2005)**
2. **Liu Gui-Mei (30 June 2005 till 31 May 2006)**

PhD students recruited & funded for this project:

1. **Chen Jin,**
   - Joined 1st October 2003
   - Passed qualifying exam in Feb 2003 at SOC
   - Nationality: China
   - Supervisors: Wynne Hsu, Lee Mong Li, Ng See-Kiong
   - Thesis committee: Not required.
2. **Hugo Willy**
   - Joined 1st October 2003
   - Passed qualifying exam in May 2004 at SOC
   - Nationality: Indonesian (Singapore PR)
   - Supervisor: Ng See-Kiong, Ken Sung
   - Thesis committee: To be formed
3. **Wong Swee Seong**
   - Joined 1st April 2004
   - Passed qualifying exam January 2004 at SOC
   - Nationality: Singapore
   - Supervisor: Ken Sung, Wong Limsoon
   - Thesis committee: To be formed

PhD students affiliated to this project but funded by other sources:

4. **Kenny Chua**
   - Joined 1st August 2003
   - AGA/NUS NGS funded
   - Passed qualifying exams at Sept 2005 at SOC
   - Supervisors: Ken Sung, Wong Limsoon
   - Thesis committee: Lee Mong Li, Ng See-Kiong
5. **Liu Huiqing (staff)**
   - Joined 1st August 2003
   - Self-funded (part-time)
   - Passed qualifying exams at SOC
   - Thesis submitted: April 2004
   - PhD awarded: December 2004
   - Supervisor: Wong Limsoon

- Thesis committee: Wynne Hsu, Li Jinyan
- Examiners: Lee Mong Li, Kwoh Chee Keong (NTU), Phil Long (Columbia), Prasanna Kolatkar (GIS)

6. **V. S. Sundararajan**
   - Joined 1st August 2003
   - I^2R funded
   - Passed qualifiying exams at SOC
   - Supervisor: Wong Limsoon
   - Thesis committee: Not set up yet
7. **Judice Koh (staff)**
   - Joined 1st August 2003
   - Self-funded (part-time)
   - Passed qualifying exams at SOC
   - Supervisor: Vladimir Brusic, Lee Mong Li
   - Thesis committee: Stephane Bressan, Sam Sung
8. **Hou Yuna**
   - Joined 1st October 2003
   - SOC funded
   - Passed qualify exam at SOC
   - Supervisor: Wynne Hsu
   - Thesis committee: Not required
9. **Li Haiquan**
   - Joined 1st August 2003
   - I^2R funded
   - Passed qualifying exams at SOC
   - Supervisor: Wee Sun Lee and Jinyan Li
   - Thesis committee: Not set up yet
10. **Feng Meng Ling**
    - Joined March 2004
    - AGA/NTU NGS funded
    - Supervisor: Tan Yap Peng (NTU), Wong Limsoon
    - Thesis committee: Not set up yet
11. **Lee Terk Shuen**
    - Joined February 2004
    - AGA/NUS NGS funded
    - Supervisors: Sam Ge, Wong Limsoon
    - Thesis committee: Jinyan Li, Wynne Hsu
12. **Donny Soh**
    - Joined December 2004
    - AIP/IC funded
    - Supervisors: Wong Limsoon, Jinyan Li, Yike Guo (IC)
    - Thesis committee: Not required.

3. Project Progress (use additional pages if necessary)

(a) The extent to which the original project objectives have been achieved.
(b) Significant changes in the research compared with the original proposal.
(c) Scheduled completion rate versus actual completion rate to date with explanation on major variances.
(d) Difficulties, if any, encountered that impeded the progress of the research and actions taken to overcome those difficulties.

Objectives:

Develop methods for analysis suitable for clinical and biological data.

Milestones (Past Period):

A. Data mining technologies **(ongoing)**

- Haiquan Li has also started participating in this project. Jinyan, Limsoon, Haiquan, Meng Ling investigated fundamental aspects of equivalence classes of patterns, as well as more sophisticated types of patterns like odds ratio patterns, relative risk patterns. Taking advantage of the convexity of pattern equivalence classes, a very fast method---gr_growth---for mining their borders was proposed and implemented. Experiments on benchmark datasets showed that gr_growth was able to produced paired key patterns and closed patterns at a speed faster than all known methods that produced only key patterns, and at a speed comparable to all known fast methods that produced only closed patterns. We also postprocessed the output of gr_growth to obtained odds ratio patterns and relative risk patterns efficiently---this was the first time it was possible to mine such sophisticated patterns efficiently.

B. Gene feature recognition **(ongoing)**

- Ken Sung, Limsoon, and Kenny Chua studied an approach---first used in Pairwise SVM---of first extracting a feature vector from a sequence where each position in the vector indicates the presence or absence of specific "domain" or signature, and then making functional assignment based on such a feature vector. In particular, we discovered that it was possible to improve considerably the accuracy of Pairwise SVM by using raw scores (as opposed to log P-values) with more relaxed gap penalties.

E. Pathway informatics **(ongoing)**

- Chen Jin, under supervision of See-Kiong, Wynne, and Mong Li, did researched on post-analysis of high-throughput protein interaction screening results for the purpose of (1) predicting the importance of a hit from such screens, and (2) detecting false positives from such screens. The "interaction pathway reliability" index proposed to rank protein-protein interactions extracted from yeast two-hybrid expts was verified by Prasanna Kolatkar of GIS to be effective. In particular, Kolatkar looked at interacting pairs that had different cellular localizations, and found that the high-scoring pairs tended to be natural cross talkers, while the low-scoring pairs tended to be not. Chen Jin made a further important observation that the same "interaction pathway reliability" index could be applied to detect false negative. This was a very unexpected bonus. Further work is ongoing to further improve the index.
- The second project initiated with Chen Jin et. al. to investigate the problem of mining for network motifs in protein-protein interaction data was in progress and did not have results to report this quarter**.**
- A third project has been initiated with Haiquan Li, Jinyan Li, and Limsoon to study the

connection between frequent pattern mining and interaction motif mining.

1. Intelligent Data Warehousing with application to Bioinformatics **(ongoing)**

- Judice, under supervision of Vladimir Brusic and Mong Li, started research on various aspects of constructing data warehouses in molecular biology. Her first project is to devise a data cleaning method for molecular data. Preliminary results include (1) A taxonomy of errors and imperfections observed in molecular databases (2) A novel method of data cleaning using association rule induction. To date some 40 database defects have been identified, and Judice has been designing an algorithm for cleaning database records. Several papers have been published or are in preparation. The latest research on this project involves automated updating of biological data warehouses.

Milestones (Reporting Period):

A. Data mining technologies **(ongoing)**

- Meng Ling has begun investigating how the frequent pattern space and equivalence classes evolve when updates are made to the underlying transaction database. He has identified a complete characterization of the evolution of frequent pattern space and equivalence classes, providing the exact conditions under which (1) a key pattern would remain key pattern, (2) a closed pattern would remain closed pattern, (3) an equivalence class would remain unchanged, (4) a key pattern would no longer be a key pattern, (5) a closed pattern would no longer be a closed pattern, (6) an equivalence class would split, and (7) two equivalence classes would merge. A paper is being prepared.
- Meng Ling has also found optimizations that lead to some performance improvement to the gr-growth and gc-growth mining algorithms developed in the last reporting period. The results were reported in PODS 2005, the top database theory conference.
- Meng Ling and Terk Shuen have also begun investigating the robustness of classifiers based on emerging patterns, odds ratio patterns, and other patterns. Intuitively, the use of multiple such patterns should make a classifier more resistant to missing or incorrect values in the data.
- Gui Mei, Jinyan, Limsoon have begun investigating more compact and succinct representation of frequent pattern spaces. One of the first ideas being considered is that of positive borders, which are frequent patterns whose subpatterns are all generators but themselves are not generators. We showed that there were much less number of such positive border patterns, compared with closed patterns, key patterns, and negative borders. We also showed that the gr-growth algorithm developed in the last reporting period could be easily adapted to mine positive borders in an efficient way. A paper is being prepared.

B. Gene feature recognition **(ongoing)**

- Ken Sung, Limsoon, and Kenny Chua discovered that about 30% of proteins share functions with their 2nd level neighbours while about 60% share functions with their immediate neighbours, whereas 3rd level neighbours are no more informative than random for the purpose of protein function prediction. Based on this observation, we have developed a method for predicting protein functions from protein interaction data. Our method outperformed a number of existing methods by a large margin. A paper is being prepared. We are further investigating the possibility of incorporating additional information besides protein interaction data for function prediction.
- Hugo Willy, under supervision of Wing-Kin Sung, See-Kiong, collaborated with Jesper Jansson to study RNA secondary structure prediction. In this project we improved the current best algorithm on inferring the secondary structure of an RNA given another known similar RNA structure. The similarity information might be based on the fact that RNAs with similar function tends to share similar structure. By using sparsification on a new recursive dynamic programming algorithm and applying a Hirschberg-like traceback technique with

compression, we obtain an improved algorithm that runs in better both in time and space complexity compared to the current best algorithm. The preliminary result of this work was published in WABI 2004 and the improved result is accepted for publication in ALGORITHMICA in September 2005.

- Liang Huai, Wynne, Mong Li, and Limsoon considered the problem of recognition of micro RNA precursors from genomic sequences. We have prepared several large files of training samples for several species. We have also made initial tests based on the "feature generation, feature selection, feature integration" approach. Our classifiers were able to achieve 70-80% sensibility and 90-97% specificity. While these numbers are comparable to existing de novo methods, they are not good enough for practical used at the whole genome level. Further investigations are needed.

E. Pathway informatics **(ongoing)**

- Chen Jin, under supervision of See-Kiong, Wynne, and Mong Li, did researched on post-analysis of high-throughput protein interaction screening results for the purpose of (1) predicting the importance of a hit from such screens, and (2) detecting false positives from such screens. We have designed a novel interaction reliability metric called "Interaction Reliability by Alternate Path" (IRAP) based on the network topological characteristics of the underlying interaction graph. A candidate interaction is considered to be reliable if it is involved in a closed loop in which the alternative path of interactions between the two interacting proteins is strong. We also devised and implemented an algorithm called AlternativePathFinder to compute the IRAP value for each interaction in a complex interaction network. Our evaluation results on real experimental yeast interaction data showed that IRAP was effective for the purpose of (1) and (2) above, and it outperformed other related network topological based approaches. Two papers (a conference paper at ICTAI and a journal paper in AIM) have been published. We are currently extending the method to also detect false negatives in the network data, as well as to devise an iterative approach to computationally purify experimental interaction data from both false positives and false negatives.
- Chen Jin, under supervision of See-Kiong, Wynne, and Mong Li, began investigation on the problem of mining for network motifs in protein-protein interaction data. An efficient algorithm called NeMoFinder has been devised and implemented for finding frequent and unique network motifs from large networks. Our preliminary results showed that using the network motifs mined from the biological networks can achieve better results in applications such as those described above than using pre-defined topological motifs (e.g. IRAP). We plan to investigate further in this, as well as extend our approach to find labeled network motifs.
- Hugo Willy and Soon-Heng Tan, under supervision of Wing-Kin Sung, See-Kiong, researched on Mining Interacting Motif from Protein Interaction Data. Recent research showed that the interactions among proteins are sometimes mediated by the short amino acid sequences of the proteins, called binding motifs. Extracting these motifs accurately and efficiently would greatly help to understand different pathways and genetic diseases. Unfortunately, biological experiments to do the latter is laborious and expensive while conventional computational approaches based on finding motifs in protein sequence set is not inadequate since they require the user to know the grouping of proteins containing the motifs from the interaction data beforehand. We propose a novel clique-motivated approach to mine the protein-protein interaction data directly to find interacting motif pairs. Experimental results show that our algorithm is able to obtain biologically significant motifs even in the presence of noise. This work is still in progress.
- Haiquan Li, Jinyan Li, and Limsoon studied the connection between frequent pattern mining and interaction motif mining. We found that every complete bi-partite subgraph of a graph is in a complete bijective correspondence to a pair of distinct closed patterns in the

adjacency matrix of the graph. The problem of enumerating complete bi-partite subgraphs is known to have complexity $O(2^a)$, where “a” is the aboricity of the graph. The correspondence to closed patterns allowed data mining algorithms to be used that are much faster than traditional graph algorithms on the average cases. This result was reported in PKDD05.

A. Intelligent Data Warehousing with application to Bioinformatics **(ongoing)**

- Judice, under the supervision of Mong Li and Vladimir, reported 11 types and 28 sub-types of biological data artifacts observed in major sequence databases, and proposed existing data cleaning solutions for some of the artifacts. Details of the analysis had been submitted in a research paper. Judice, Mong Li, and and Wynne also devised a new correlation-based method to detect attribute outliers. The method could be applied to identify annotation errors in sequence databases. The method paper had been submitted. Further improvements to the method and its application are in progress.

4. Milestones / Deliverables

(a) Performance Indicators

Note: Please do not change the table. If an item is not applicable to your project, indicate with "N/A".

Economic Impact Indicators

		Actual for reporting period - incremental	Cum actual for project	Cum target for project
Attracting investments	Total value of investments attracted to Singapore for which the project plays a significant role	0	0	
Improving the competitiveness of local industry	Cash contribution from industry	0	0	
	In-kind contribution from industry	0	0	
	Royalties and licensing fees from IP	0	0	
	No. of joint projects with industry	0	0	
	Value (total project cost) of joint projects with industry	0	0	
Creating new industries	No. of spin-off companies or joint ventures (please provide listing)	0	0	
	No. of licensing agreements	0	0	
	No. of new products or processes developed	0	0	
	No. of new products or processes commercialised	0	0	

Capability Building Indicators

Carrying out world class research	No. of patents filed (please provide listing)	0	0	
	No. of patents granted (please provide listing)	0	0	
	No. of papers published in prestigious journals or conferences (please provide listing)	23 (includes: 1 book and 5 book chaps)	81 (include: 6 books, 14 book chaps)	
	No. of joint programmes with prestigious international research organisations	0	0	
Developing manpower	No. of undergraduate/polytechnic students attached to the project for more than 3 months	0	0	
	No. of postgraduate students attached to the projects for more than 3 months	0	11	
	No. of RSEs trained through formal post-graduate studies	0	1	
	No. of RSEs trained through joint projects	0	1	
	No. of conferences & seminars	1 (4th joint	6	

	organised (please provide listing)	lab mtg/seminar)		

(b) Additional milestones/deliverables not captured in 4(a) above (use additional pages if necessary)

Conferences & seminars organized

New:

1. 16/8/05. Internal seminars: “Function Prediction from Protein Interactions” by Kenny Chua, and “Accurate Identification of MicroRNA Precursors” by Yang Lianghuai.

.

Cumulative:

2. 17-21/1/05. 3rd Asia-Pacific Bioinformatics Conference, held at I^{2}R, with Vladimir Brusic as organizing chair. The conference attracted 118 full paper submissions, with 35 selected for plenary presentations, and attracted over 200 participants from 22 countries.
3. 24/1/05. I2R-SOC Joint Meeting with Yuan Ze Univ Taiwan, held at I^{2}R, with Ng See-Kiong as organizing chair. Students in our joint lab made 8 short presentations on our work to our Taiwanese visitors lead by Prof. Shu-Yuan Chen.
4. 28/11/03. Internal seminar: “On Combining Multiple Microarray Studies for Improved Functional Classification by Whole-Dataset Feature Selection” by V. S. Sundararajan.
5. 28/11/03. Internal seminar: “Efficient Remote Homolog Detection Using Local Structure” by Hou Yuna
6. 30/9/03. Internal seminar: “Belief Propagation with Dynamic Arc Removal for Systematic Assessment of High-Throughput Protein Interaction Data” by Chen Jin.

Patents

New:

1.

Cumulative:

2.

Papers

New:

1. Guozhu Dong, Jinyan Li, Limsoon Wong. **The Use of Emerging Patterns in the Analysis of Gene Expression Profiles for the Diagnosis and Understanding of Diseases**. *New Generation of Data Mining Applications*, edited by M. Kantardzic and J. Zurada, chapter 14, pages 331--354. John Wiley, April 2005. (book chapter)
2. S.-K. Ng. **Constructing Biological Networks of Protein-Protein Interactions**. Information Processing and Living Systems, edited by V.B. Bajic and T.W. Tan, chapter 7, pages 653-669. Imperial College Press, 2005 (book chapter)
3. T.W. Tan, K.H. Choo, J.C. Tong, M.T. Tammi, V.B. Bajic, **Biological Databases and Web Services: Metrics for Quality**, *Information Processing and Living Systems*, edited by V.B. Bajic and T.W. Tan, chapter 16, pages 771-777. Imperial College Press, 2005. (book chapter)
4. K. Rajaraman, L. Zuo, V. Choudhary, Z. Zhang, V.B. Bajic, H. Pan, T.-S. Sim, S. Swarup, **Overview of text-mining in life sicences**, *Information Processing and Living Systems*, edited by V.B. Bajic and T.W. Tan, chapter 9, pages 687-694. Imperial College Press, 2005. (book chapter)
5. E. Huang, L. Yang, R. Chowdhary, A. Kassim, V.B. Bajic, **An algorithm for ab initio**

DNA motif detection, *Information Processing and Living Systems*, edited by V.B. Bajic and T.W. Tan, chapter 4, pages 611-614. Imperial College Press, 2005. (book chapter)

6. V.B. Bajic, T.W. Tan (Eds.). *Information Processing and Living Systems*, Imperial College Press, 2005. (book)
7. Huiqing Liu, Hao Han, Jinyan Li, Limsoon Wong. **DNAFSMiner: A Web-Based Software Toolbox to Recognize Two Types of Functional Sites in DNA Sequences.** *Bioinformatics*, 21:671--673, March 2005.
8. Xu Peng, R. Krishna Murthy Karuturi, Lance D. Miller, Kui Lin, Yonghui Jia, Pinar Konda, Long Wang, Limsoon Wong, Edison T. Liu, Mohan K. Balasubramanian, Jianhua Liu. **Identification of Cell Cycle-regulated Genes in Fission Yeast**. *Molecular Biology of the Cell*, 16(3):1026--1042, March 2005.
9. Huiqing Liu, Jinyan Li, Limsoon Wong. **Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data**. *Bioinformatics*, 21(16):3377--3384, 2005.
10. Guozhu Dong, Chunyu Jiang, Jian Pei, Jinyan Li, Limsoon Wong. **Mining Succint Systems of Minimal Generators of Formal Concepts**. *Proceedings of 10th International Conference on Database Systems for Advanced Applications*, pages 175--187, Beijing, China, April 2005.
11. Li Lin, Limsoon Wong, Tzeyun Leong, Pohsan Lai. **LinkageTracker: A Discriminative Pattern Tracking Approach to Linkage Disequilibrium Mapping**. *Proceedings of 10th International Conference on Database Systems for Advanced Applications*, pages 30--42, Beijing, China, April 2005.
12. Haiquan Li, Jinyan Li, Limsoon Wong, Mengling Feng, Yap-Peng Tan. **Relative Risk and Odds Ratio: A Data Mining Perspective**. *Proceedings of 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 368--377, Baltimore, Maryland, June 2005.
13. Jinyan Li, Haiquan Li, Donny Soh, Limsoon Wong. **A Correspondence Between Maximal Complete Bipartite Subgraphs and Closed Patterns**. *Proceedings of 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 146--156, Porto, Portugal, October 2005.
14. X.-L. Li, S.-H. Tan, S.-K. Ng. **Exploring Cross-Function Domain Interaction Map**, Proceedings of *2005 International Joint Conference of InCoB, AASBi and KSBI (BIOINFO 2005)*, Busan, Korea, September 22-24, 2005, pages 431-436.
15. X-L. Li, S.-H. Tan, S.-K. Ng. **Protein Interaction Prediction using Inferred Domain Interactions and Biologically-Significant Negative Dataset**, Proceedings of *International Conference on Computational Science and Its Applications (ICCSA-2005)*, Apr 2005, Pages 318 - 326.
16. Choy, J. Jansson, K. Sadakane, and W. K. Sung. **Computing the Maximum Agreement of Phylogenetic Networks**. *Theoretical Computer Science*, 335(1): 93-107, 2005.

17. W. Leong, F. P. Preparata, W. K. Sung, and H. Willy. **On the control of hybridization noise in DNA Sequencing-by-Hybridization**. *Journal of Bioinformatics and Computational Biology*, 3(1):79-98, 2005.

18. Arunkumar, W. K. Sung, and A. Mittal. **Protein structure and fold prediction using tree-augmented naive Bayesian classifier**. *Journal of Bioinformatics and Computational Biology*, 3(4):803-820, 2005.

19. 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.

20. Partick Ng, Chia-Lin Wei, Wing-Kin Sung, Kuo Ping Chiu, Leonard Lipovich, Chin Chin Ang, Sanjay Gupta, Atif Shahab, Azmi Ridwan, Chee Hong Wong, Edison T. Liu, Yijun Ruan. **Gene Identification Signature (GIS) Analysis for Transcriptome Characterization and Genome Annotation**. *Nature Methods*, 2:105-111, 2005.

21. J. Wang, W. K. Sung, A. Krishnan, and K. B. Li. **Protein subcellular localization prediction for Gramnegative bacteria using amino acid subalphabets and a combination of multiple support vector machines**. *BMC Bioinformatics*, 6:174, 2005.

22. V. Narang, W. K. Sung, and A. Mittal. **Computational Modeling of Oligonucleotide Positional Densities for Human Promoter Prediction**. *Artificial Intelligence in Medicine*, 35:107-119, 2005.

23. Trinh N. D. Huynh, J. Jansson, N. B. Nguyen, and W. K. Sung. **Constructing a Smallest Refining Galled Phylogenetic Network**. In *RECOMB*, 265-280, 2005.

Cumulative:

24. Louxin Zhang, Limsoon Wong, editors. *Selected Topics in Post-Genome Knowledge Discovery*, Singapore University Press, Singapore, May 2004. (book)
25. Limsoon Wong, editor. *The Practical Bioinformatician*, World Scientific, New Jersey, May 2004. (book)
26. Yi-Ping Phoebe Chen, Limsoon Wong, editors. *Proceedings of the 3rd Asia-Pacific Bioinformatics Conference*, Imperial College Press, London, January 2005. (book)
27. T. Akutsu, V. Brusic, M. Kanehisa, S. Miyano, T. Takagi, editors. *Proceedings of 15th International Conference on Genome Informatics*, Universal Academy Press, Tokyo, Japan, December 2004. (book)
28. V. Brusic, A.M. Khan, editors. *Abstract Book of the 3rd Asia-Pacific Bioinformatics Conference & Singapore Bioinformatics Week*. World Scientific, Jan. 2005. (book)
29. Mohammed Zaki, Limsoon Wong. **Data Mining Techniques**. *Selected Topics in Post-Genome Knowledge Discovery*, edited by Limsoon Wong, Louxin Zhang, chapter 4, pages 125--163. National University of Singapore Press, May 2004. (book chapter)
30. Jinyan Li, Huiqing Liu, Anthony Tung, Limsoon Wong. **Data Mining Techniques for the Practical Bioinformatician**. *The Practical Bioinformatician*, edited by Limsoon Wong, chapter 3, pages 35-70, World Scientific, May 2004. (book chapter)
31. Jinyan Li, Huiqing Liu, Limsoon Wong, Roland Yap. **Techniques for Recognition of Translation Initiation Sites.** *The Practical Bioinformatician*, edited by Limsoon Wong, chapter 4, pages 71-90, World Scientific, May 2004. (book chapter)
32. Jinyan Li, Limsoon Wong. **Techniques for Analysis of Gene Expression.** *The Practical Bioinformatician*, edited by Limsoon Wong, chapter 14, pages 319-346, World Scientific, May 2004. (book chapter)
33. Limsoon Wong. **Technologies for Biological Data Integration.** *The Practical Bioinformatician*, edited by Limsoon Wong, chapter 17, pages 375-400, World Scientific, May 2004. (book chapter)
34. Kui Lin, Jianhua Liu, Lance Miller, Limsoon Wong. **Genome-Wide cDNA Oligo Probe Design and its Applications in *Schizosaccharomyces Pombe*.** *The Practical Bioinformatician*, edited by Limsoon Wong, chapter 15, pages 347-358, World Scientific, May 2004. (book chapter)
35. See-Kiong Ng. **Molecular Biology for the Practical Bioinformatician**. *The Practical Bioinformatician*, edited by Limsoon Wong, Chapter 1, pages 1-30, World Scientific, May 2004. (book chapter)
36. Soon Heng Tan, See-Kiong Ng. **Discovering Protein-Protein Interactions**. *The Practical*

Bioinformatician, edited by Limsoon Wong, Chapter 13, pages 293-318, World Scientific, May 2004. (book chapter)

37. Wing Kin Sung. **RNA Secondary Structure Prediction**, *The Practical Bioinformatician,* edited by Limsoon Wong, Chapter 8, pages 167-192, , World Scientific, May 2004. (book chapter)
38. Huiqing Liu, Jinyan Li, Limsoon Wong**. Selection of Patient Samples and Genes for Outcome Prediction**. *IEEE Bioinformatics Proceedings (CSB2004),* pages 382--392, Stanford, CA, August 2004.
39. Jinyan Li, Hwee-Leng Ong. **Feature Space Transformation for Better Understanding Bio-Medical Classifications.** *Journal of Research and Practice in Information Technology*. 36(3): 131-144, August 2004.
40. Shao-Wu Meng, Zhuo Zhang, Jinyan Li. **Twelve C2H2 Zinc Finger Genes on Human Chromesone 19 can Be Each Translated into the Same Type of Protein after Frameshifts.** *Bioinformatics*. 20(1): 1-4, 2004.
41. Jinyan Li, Thomas Manoukian, Guozhu Dong, Kotagiri Ramamohanarao**. Incremental Maintenance on the Border of the Space of Emerging Patterns.** *Data Mining and Knowledge Discovery*. Volume 9, issue 1, pages 89-116, 2004.
42. Jinyan Li, Guozhu Dong, Kotagiri Ramamohanarao, Limsoon Wong**. DeEPs: A New Instance-based Discovery and Classification System.** *Machine Learning*. 54(2): 99--124, 2004.
43. Jinyan Li, Kotagiri Ramamohanarao. **A Tree-based Approach to the Discovery of Diagnostic Biomarkers for Ovarian Cancer.** *Proceedings of the Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2004), Sydney, Australia. May 26-28, 2004*. Pages: 682-691.
44. Soon Heng Tan, Zhuo Zhang and See-Kiong Ng**. ADVICE: Automated Detection and Validation of Interaction by Co-Evolution**, *Nucleic Acids Research*, 32:W69-W72, 2004.
45. Zhuo Zhang and See-Kiong Ng. **InterWeaver: Interaction Reports for Discovering Potential Protein Interaction Partners with Online Evidence**, *Nucleic Acids Research,* 32:W73-W75, 2004.
46. Z. Zhuo, S. Tang, S.-K. Ng. **Toward Discovering Disease-Specific Gene Networks**, *Proceedings of 3rd Asia-Pacific Bioinformatics Conference,* Singapore, pages 161-170, 17-21 January, 2005.
47. J. Chen, W. Hsu, M.L. Lee, S.-K. Ng. **Systematic Assessment of High-Throughput Experimental Data for Reliable Protein Interactions using Network Topology**. *Proceedings of 16th IEEE International Conference on Tools with Artificial Intelligence,* Florida, pages 368-372, 15-17 November, 2004,.
48. J. Jansson, S.-K. Ng, W.-K. Sung, H. Willy. **A Faster and More Space-Efficient Algorithm for Inferring Arc-Annotations for RNA Sequences through Alignment**, *Proceedings of 4th Workshop on Algorithgms in Bioinformatics*, Bergen, Norway, pages 302-313, 14-17 September, 2004.
49. S.-H. Tan, W.-K. Sung and S.-K. Ng. **An Automated Approach for Protein Motif Discovery Using Interaction-Driven Motif Mining**, *Proceedings of 2nd International Conference on Computer Science and Its Applications*, San Diego, pages 224-232, 28-30 June, 2004.
50. S.-H. Tan, W.-K. Sung, S.-K. Ng. **Discovering Novel Interacting Motif Pairs from Large Protein-Protein Interaction Datasets**, *Proceedings of 4th IEEE Symposium of Bioinformatics and Bioengineering*, pages 568-575, 19-21 May, 2004.
51. Koh J.L.Y. and Brusic V. **Warehousing of Biological Data**. *International Workshop on Knowledge Discovery in BioMedicine (KDbM-04).* A PRICAI 2004 Workshop, August 2004
52. Koh J.L.Y., Lee M.L., Khan A.M., Tan P.T.J and Brusic V**. Duplicate Detection in**

Biological Data using Association Rule Mining. *2nd European Workshop on Data Mining and Text Mining for Bioinformatics*. A ECML/PKDD 2004 workshop, Pisa, Italy, September 24, 2004.

53. Koh J.L.Y., Krishnan S.P.T., Seah S.H., Tan P.T., Khan A., Lee M.L. and Brusic V. (2004). **BioWare: A framework for bioinformatics data retrieval, annotation and publishing**. *Search and Discovery in Bioinformatics*. A SIGIR 2004 Workshop, July 29, 2004, Sheffield, UK.
54. Koh J.L.Y. and Brusic V. **Bioinformatics Database Warehousing**. In Chen YPP, (ed.) *Bioinformatics Technology*. Springer, pages 45-62, Jan. 2005.
55. 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.
56. Judice L.Y. Koh, Mong Li Lee, Vladimir Brusic. **A Classification of Biological Data Artifacts**, in *ICDT Workshop on Database Issues in Biological Databases (DBiBD),* Edinburgh, Scotland, UK, January 2005.
57. Yuna Hou, Wynne Hsu, Mong Li Lee, Chris Bystroff. **Remote Homolog Detection Using Local Sequence-Structure Correlations**, *PROTEINS: Structure, Function, and Bioinformatics*, 57(3):518-530, Nov 2004.
58. Yuna Hou, Wynne Hsu, Mong Li Lee, Chris Bystroff**. Efficient Remote Homology Detection Using Local Structure**, *Bioinformatics*, 19(17):2294-2301, Nov 2003.
59. Hon Nian Chua and Wing-Kin Sung. **A better gap penalty for pairwise SVM**. *Proceedings of 3rd Asia-Pacific Bioinformatics Conference,* Singapore, pages 11-20, 17-21 January, 2005
60. Y.-J. He, T.N.D. Huynh, J. Jansson, W.-K. Sung**. Inferring phylogenetic relationships avoiding forbidden rooted triplets.** *Proceedings of 3rd Asia-Pacific Bioinformatics Conference,* Singapore, pages 339-348, 17-21 January 2005
61. Rajesh Chowdhary, R. Ayesha Ali, Vladimir B. Bajic. **Modeling 5' regions of histone genes using Bayesian networks**. **.** *Proceedings of 3rd Asia-Pacific Bioinformatics Conference,* Singapore, pages 283-288, 17-21 January 2005.
62. W. K. Hon, M. Y. Kao, T. W. Lam, W. K. Sung and S. M. Yiu. **Non-shared Edges and Nearest Neighbor Interchange Revisited**. *Information Processing Letters*, 91(3):129-134, 2004.
63. Wing-Kai Hon, Ming-Yang Kao, Tak-Wah Lam, Wing-Kin Sung and Siu-Ming Yiu. **Subtree Transfer Distance for Degree-D Phylogenies**. *International Journal of Foundations of Computer Science* (IJFCS), 15(6):893-909, 2004.
64. J. Jansson, C. Choy, K. Sadakane, and W. K. Sung**. Computing the Maximum Agreement of Phylogenetic Networks**. *Theoretical Computer Science,* to appear.
65. Jesper Jansson, Trung Hieu Ngo, and Wing-Kin Sung. **Local Gapped Subforest Alignment and Its Application in Finding RNA Structural Motifs**. *Proceedings of ISAAC*, pages, 569-580, 2004.
66. Tie-Fei Liu, Wing-Kin Sung and Ankush Mittal. **Learning Multi-Time Delay Gene Network Using Bayesian Network Framework**. *Proceedings of 16th IEEE International Conference on Tools with Artificial Intelligence* (ICTAI), pages 640-64, 2004.
67. Ravi Gupta, Ankush Mittal, Wing-Kin Sung and Vipin Narang**. Detection of Palindromes in DNA sequences using Periodicity Transform**. *Proceedings of IEEE International Workshop on BioMedical Circuits and Systems*, 2004.
68. Huiqing Liu, Hao Han, Jinyan Li, Limsoon Wong**. Using Amino Acid Patterns to Accurately Predict Translation Initiation Sites.** *In Silico Biology*, 4(3):255-269, 2004.
69. See-Kiong Ng, Limsoon Wong**. Accomplishments and Challenges in Bioinformatics**. *IEEE IT Professional Magazine*, 6(1):44-50, January/February 2004.
70. Huiqing Liu, Hao Han, Jinyan Li, Limsoon Wong. **An in silico method for prediction of**

polyadenylation signals in human sequences. *Proceedings of 14th International Conference on Genome Informatics* , pages 84--93, Yokohama, December 2003.

71. Jinyan Li, Huiqing Liu, Limsoon Wong**. Use of Built-in Features in the Interpretation of High-dimensional Cancer Diagnosis Data**. *Proceedings of 2nd Asia Pacific Bioinformatics Conference*, pages 67--74, Dunedin, New Zealand, January 18-22, 2004.
72. See-Kiong Ng, Soon-Heng Tan, V. S. Sundararajan**. On Combining Multiple Microarray Studies for Improved Functional Classification by Whole-Dataset Feature Selection.** *Proceedings of 14th International Conference on Genome Informatics*, pages 44--53, Yokohama, December 2003.
73. Jin Chen, Wynne Hsu, Mong Li Lee**. Order-Sensitive Clustering for Remote Homologous Protein Detection**, *Proceedings of 15th IEEE International Conference on Tools with Artificial Intelligence*, Sacramento, California, November 2003.
74. HaiQuan Li, Jinyan Li, Soon-Heng Tan, See-Kiong Ng**. Binding Motif Pair Discovery from Protein Complex Structural Data and Protein Interaction Sequence Data.** *Proceedings of 9th Pacific Symposium for Biocomputing*, pages 312-323, January 2004.
75. See-Kiong Ng, Zhexuan Zhu, Yew-Soon Ong. **Whole-Genome Functional Classification of Genes by Latent Semantic Analysis on Microarray Data**. *Proceedings of 2nd Asia-Pacific Bioinformatics Conference*, pages 123-129, New Zealand, January 2004.
76. See-Kiong Ng, Soon-Heng Tan**. Discovering Protein-Protein Interactions.** *Journal for Bioinformatics and Computational Biology*, 1(4):711-741, 2004.
77. T. F. Liu, P. L. Mao, and A. Mittal , W.-K. Sung**, Tag SNP selection by brute-force and heuristic algorithms,** *Proceedings of German Conference on Bioinformatics* (GCB), 2003.
78. Wei Chen , W.-K. Sung, **On Half Gapped Seeds**. *Proceedings of 14th International Conference on Genome Informatics*, Yokohama, December 2003.
79. Chinnasamy Arunkumar, Ankush Mittal, W.-K. Sung, **Protein Structure and Fold Prediction Using Tree-Augmented Bayesian Classifier**. *Proceedings of Pacific Symposium on Biocomputing* (PSB), 2004.
80. Charles Choy, Jesper Jansson, and Kunihiko Sadakane, W.-K. Sung**. Computing the Maximum Agreement of Phylogenetic Network**. Proceedings of 10th Australasian Theory Symposium (CATS), 2004.
81. Jesper Jansson, Kunihiko Sadakane, and Ng Hon Keong, W.-K. Sung**. Rooted Maximum Agreement Supertrees.** *Proceedings of Latin American Theoretical Informatics (LATIN)*, pages 499—508, 2004.

Invited talks

New:

1. Jinyan Li, Limsoon Wong. **Bioinformatics and Machine Learning**. *Tutorial talk at 20th National Conference on Artificial Intelligence (AAAI-05)*, Pittsburgh, 9-13 July 2005.
2. Ken Sung, Limsoon Wong. **Bioinformatics in Practice**. *Invited tutorial at 8th International Conference on Discovery Science*, Marina Mandarin Hotel, Singapore, 8 October 2005.
3. Limsoon Wong. **Selection of Patient Samples and Genes for Disease Prognosis**. *Invited talk at "Informatics Inspired Biology" Symposium*, BioPolis, Singapore, 16 January 2005.
4. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments.** *Invited keynote at 3rd Korea-Singapore Joint Workshop on Bioinformatics and Natural Language Processing*, Muju Resort, 20-22 February 2005.
5. Limsoon Wong. **Building gene networks by information extraction, cleansing, and integration**. *Invited plenary lecture at 3rd International Symposium on e-Biology Initiative: Towards New Frontiers of Biology*, Takeda Hall, University of Tokyo, Tokyo, Japan, 11 March 2005.
6. Limsoon Wong. **Some interesting issues in constructing gene/protein networks**, *Invited round-table presentation at 3rd International Symposium on e-Biology Initiative: Towards*

New Frontiers of Biology, National Institute of Informatics, Tokyo, Japan, 10 March 2005.

7. Haiquan Li, Jinyan Li, Limsoon Wong. **Binding Motif Pairs from Interacting Protein Groups**. *Invited talk at Workshop on Data Analysis and Data Mining in Proteomics*, Institute for Mathematical Sciences, NUS, Singapore, 12 May 2005.
8. Limsoon Wong. **Building Gene Networks by Information Extraction, Cleansing, & Integration.** *Invited talk at Tamkang University*, Taiwan, 31 May 2005.
9. Limsoon Wong. **Discovering Binding Motif Pairs from Interacting Protein Groups.** *Invited talk at Tamkang University*, Taiwan, 31 May 2005.
10. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments.** *Invited talk at Tamkang University*, Taiwan, 1 June 2005.
11. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments.** *Invited talk at Changchun International Bioinformatics Workshop*, Changchun, Jilin, China, 5-7 July 2005.
12. Hon-Nian Chua, Wing-Kin Sung, Limsoon Wong. **Protein Function Prediction from Protein Interactions.** *Invited talk at Singapore-Bangalore Biomedical Symposium*, Biopolis, Singapore, 8-9 September 2005.
13. Wing-Kin Sung. **DTSeq: Decision Tree Based De Novo Peptide Sequencing**. *Invited talk at Workshop on Data Analysis and Data Mining in Proteomics*, Institute for Mathematical Sciences, NUS, Singapore, 12 May 2005.
14. See-Kiong Ng. **Mining Motifs from Protein Interaction Data**. *Invited talk at Workshop on Data Analysis and Data Mining in Proteomics*, Institute for Mathematical Sciences, NUS, Singapore, 12 May 2005.
15. See-Kiong Ng. **Computational Purification of Protein Interactomes Using Network Topology.** Invited talk at *3rd Symposium of Association of Asian Societies for Bioinformatics* (AASBi 2005), September 23 2005, Busan, Korea.

Cumulative:

16. Limsoon Wong**. Imagination to Reality: Life as a Researcher**. *Invited talk at Hwa Chong Junior College Annual Student Research Symposium,* Hwa Chong Junior College, 3 April 2004.
17. Limsoon Wong**. Exciting Media.** *Invited talk at World Scientific*, 6 April 2004.
18. Limsoon Wong. **A Practical Introduction to Bioinformatics**. *Invited short course at National Yang Ming University*, Taipei, Taiwan, 22 May 2004.
19. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments**. *Invited talk at National Yang Ming University*, Taipei, Taiwan, 26 May 2004.
20. Limsoon Wong. **Convexity of Itemset Spaces**. *Invited talk at Academia Sinica*, Taipei, Taiwan, 28 May 2004.
21. Limsoon Wong. **Diagnosis of Childhood Acute Lymphoblastic Leukaemia and Optimization of Risk-Benefit Ratio of Therapy.** *Invited talk at National Cheng Kung University*, Tainan, Taiwan, 21 May 2004.
22. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments**. *Invited talk at Genome Institute of Singapore*, 4 June 2004.
23. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments.** *Invited talk at 3rd International Conference on Bioinformatics*, Auckland, New Zealand, 4-8 September 2004.
24. Jinyan Li, Limsoon Wong. **Rule-Based Data Mining Methods for Classification Problems in Biomedical Domains.** *Tutorial given at 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Pisa, Italy, 20-24 September 2004.
25. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments.** *Invited talk at 5th HUGO Pacific Meeting and 6th Asia-Pacific Meeting on Human Genetics*, BioPolis, Singapore, 17-20 November 2004.

26. Limsoon Wong. **Gene Finding and Gene Feature Recognition by Computational Analysis**. *Invited tutorial at Beijing Normal University*, Beijing, 21-25 November 2004.
27. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments.** *Invited talk at Tsing Hua University*, Beijing, 25 November 2004.
28. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments.** *Invited talk at Edinburgh University*, Edinburgh, 1 October 2004.
29. Limsoon Wong. **Diagnosis of Childhood Acute Lymphoblastic Leukaemia and Optimization of Risk-Benefit Ratio of Therapy**. *Invited talk at Edinburgh University*, Edinburgh, 1 October 2004.
30. Limsoon Wong. **Knowledge Discovery in Biomedicine**. *Invited talk at National Healthcare Group Annual Scientific Congress*, Raffles Convention Centre, Singapore, October 2004.
31. Limsoon Wong**. Research & Discovery: Technologies Today for Solving Problems Tomorrow .** *Talk at Pre-Horizon Seminar, Infocomm Horizon 2004*, I^{2}R, Singapore, 1 November 2004.
32. Limsoon Wong**. Selection of Patient Samples and Genes for Disease Prognosis.** *Invited talk at "Informatics Inspired Biology" Symposium*, BioPolis, Singapore, 16 January 2005.
33. Vladimir Brusic**. Databases and Warehouses for Bioinformatics**. *Invited lecturer at "New Zealand Summer School of Bioinformatics"*, Knowledge Engineering and Discovery Research institute, Auckland, New Zealand, February 2004.
34. Sin Lam Tan, Vidhu Choudhary, Alan Christoffels, Byrappa Venkatesh, Vladimir B. Bajic. **Comparison of core promoters in Fugu rubripes and human**. *Invited keynote at 3rd Asia-Pacific Bioinformatics Conference*, Singapore, January 2005. (Talk given Vladimir Bajic)
35. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments.** *Invited talk at Lilly Systems Biology Symposium*, BioPolis, Singapore, 4 February 2004.
36. See-Kiong Ng. **Whole-Genome Functional Classification of Genes by Latent Semantic Analysis on Microarray Data**. Invited talk at *2nd Korea-Singapore Joint Workshop on NLP and Bioinformatics*, KAIST, Korea, 19 February 2004.
37. See-Kiong Ng. **Combining Multiple Microarray Studies for Improved Functional Classification by Whole-Dataset Feature Selection**. Invited talk at *2nd Korea-Singapore Joint Workshop on NLP and Bioinformatics*, KAIST, Korea, 19 February 2004.

Awards

New:

1. Haiquan Li. SOC Dean's Graduate Award, January 2005
2. Haiquan Li, NUS President's Graduate Fellowship, August 2005.

MILESTONES AND STATUS

A. Data mining technologies (ongoing)
B. Gene feature recognition (ongoing)
C. Gene expression analysis
D. Venome informatics (abandoned, as Vladimir Brusic is not longer working full time)
E. Pathway informatics (ongoing)
F. Intelligent Data Warehousing with application to Bioinformatics (ongoing)

Principal Investigator:

Name: _Wynne Hsu_________ (SOC)

Signature: ____________________________________ Date: _____________________

Name: __Limsoon Wong______ (I^2R/SOC)

Signature: _ __________________________ Date: 15 October 2005_

Section B: ***To be completed by I^2R Reviewing Officer:***

Comments:

Name:

Signature: ____________________________________ Date: ________________________