QUARTERLY PROGRESS REPORT

FOR I²R-SOC JOINT LAB

Section A: *To be completed by Principal Investigator*

Reporting Period	1 April 2004 to 31 January 2005
Project Vote No. at SOC	R-252-000-172-593
Title of Project	Knowledge Discovery from Biological & Clinical Data
PIs	A/Prof Wynne Hsu, SOC Prof Wong Lim Soon, I²R
Co-PIs	Dr. Lee Mong Li, SOC Dr. Ken Sung, SOC Prof. Vladimir Bajic, I²R Prof. Vladimir Brusic, I²R A/Prof. Ng See Kiong, I²R Dr. Li Jinyan, I²R
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)
EOM	$822,000	NA	$112,479.54
OOE	$45,000	NA	$11,682.74
EQPT	$70,000	NA	-
Total	$937,000	NA	$124,162.28

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. Together with the committed $174,548, the manpower utilization is about 35%. We intend to be more aggressive in our recruitment for the coming semester.

SoC has been very generous in providing high-end PCs to the research scholars. With this, we are in the process of evaluating and purchasing servers to provide a central repository for biomedical datasets and source codes for the students.

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)

Postdoc recruited & funded for this project:
1. **Yang Liang-Huai**

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 set up yet
2. **Hugo Willy**
    - Joined 1st October 2003
    - Nationality: Indonesian (Singapore PR)
    - Supervisor: Ng See-Kiong, Ken Sung
    - Thesis committee: Not set up yet
3. **Wong Swee Seong**
    - Joined 1st April 2004
    - Passed qualifying exam January 2004
    - Nationality: Singapore
    - Supervisor: Ken Sung, Wong Limsoon
    - Thesis committee: Not set up yet

PhD students affiliated to this project but funded by other sources:
4. **Kenny Chua**
    - Joined 1st August 2003
    - AGA/NUS NGS funded
    - 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, Li Jinyan
    - 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
   - Supervisor: Wynne Hsu
   - Thesis committee: Not set up yet

9. **Li Haiquan**
   - Joined 1st August 2003
   - I^{2}R 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 (previous report period):
A. Data mining technologies **(started)**
Jinyan and Limsoon has begun educating Feng Meng Ting and Lee Terk Shuen on basics of data mining. Some work has also started on analyzing the decomposition of frequent itemsets and emerging patterns and their concise representation. Some preliminary results were obtained and implemented by internship student Julian Lorenz wrt (1) enumerate all frequent patterns fast, (2) determine the support level of frequent patterns fast, (3) enumerate all patterns within a specified range of support values fast.
B. Gene feature recognition **(ongoing)**

- Kenny Chua, under supervision of Ken and Limsoon, started research on post-analysis of sequence alignment search for the purpose of (1) detecting annotation errors in sequence database, (2) inferring active site/domains, and (3) function assignment. An initial project was completed that showed ability to perform (1) above, now awaiting writing up. A second project was started to tackle (3) by using the approach 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.

E. Pathway informatics **(ongoing)**

- Chen Jin, under supervision of See-Kiong, Wynne, and Mong Li, started research 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. An initial project was completed where a "interaction pathway reliability" index was proposed to rank protein-protein interactions extracted from yeast two-hybrid expts. Preliminary analysis showed that this index was more effective than the existing "interaction generality" index. This idea was shared with Dr. Prasanna Kolarkat of GIS. Kolatkar subsequently agreed to help verify this further against additional biological knowledge at GIS over the next few month.

- A second project has been initiated with Chen Jin et. al. to investigate the problem of mining for network motifs in protein-protein interaction data.

F. 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

Milestones (report 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.

F. 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.

---

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	0	0	

	with industry			
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)	44 (includes: 5 books, 9 book chapters)	58 (include: 5 books, 9 book chapters)	
	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	1 (Donny Soh)	10	
	No. of RSEs trained through formal post-graduate studies	0	1 (Huiqin)	
	No. of RSEs trained through joint projects	0	1 (Han Hao)	
	No. of conferences & seminars organised (please provide listing)	2 (APBC2005, Joint Mtg w/ Yuan Ze Univ)	5	

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

Conferences & seminars organized
New:
1. 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.
2. 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.

Cumulative:
3. 28/11/03. Internal seminar: "On Combining Multiple Microarray Studies for Improved Functional Classification by Whole-Dataset Feature Selection" by V. S. Sundararajan.
4. 28/11/03. Internal seminar: "Efficient Remote Homolog Detection Using Local

Structure" by Hou Yuna

5. 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. Louxin Zhang, Limsoon Wong, editors. *Selected Topics in Post-Genome Knowledge Discovery*, Singapore University Press, Singapore, May 2004. (book)
2. Limsoon Wong, editor. *The Practical Bioinformatician*, World Scientific, New Jersey, May 2004. (book)
3. Yi-Ping Phoebe Chen, Limsoon Wong, editors. *Proceedings of the 3rd Asia-Pacific Bioinformatics Conference*, Imperial College Press, London, January 2005. (book)
4. 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)
5. V. Brusic, A.M. Khan, editors. *Abstract Book of the 3rd Asia-Pacific Bioinformatics Conference & Singapore Bioinformatics Week*. World Scientific, Jan. 2005. (book)
6. 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)
7. 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)
8. 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)
9. 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)
10. 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)
11. 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)
12. 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)
13. 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)
14. 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)

15. 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.

16. 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.

17. 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.

18. 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.

19. Jinyan Li, Guozhu Dong, Kotagiri Ramamohanarao, Limsoon Wong**. DeEPs: A New Instance-based Discovery and Classification System.** *Machine Learning*. 54(2): 99--124, 2004.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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,.

25. 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.

26. 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.

27. 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.

28. 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

29. 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.

30. 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.

31. Koh J.L.Y. and Brusic V. **Bioinformatics Database Warehousing**. In Chen YPP, (ed.) *Bioinformatics Technology*. Springer, pages 45-62, Jan. 2005.

32. Chen Jin, Wynne Hsu, Mong Li Lee, See-Kiong Ng. **Discovering Reliable Protein Interactions from High-Throughput Experimental Data using Network Topology.** *Artificial Intelligence in Medicine* (AIIM Special Issue on Computational Intelligence Techniques in Bioinformatics), 2005. (accepted)

33. 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.

34. 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.

35. Yuna Hou, Wynne Hsu, Mong Li Lee, Chris Bystroff**. Efficient Remote Homology Detection Using Local Structure**, *Bioinformatics*, 19(17):2294-2301, Nov 2003.

36. 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

37. 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

38. 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.

39. 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.

40. 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.

41. J. Jansson, C. Choy, K. Sadakane, and W. K. Sung**. Computing the Maximum Agreement of Phylogenetic Networks**. *Theoretical Computer Science*, to appear.

42. 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.

43. 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.

44. 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

Cumulative:

45. 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.

46. See-Kiong Ng, Limsoon Wong**. Accomplishments and Challenges in Bioinformatics**.

IEEE IT Professional Magazine, 6(1):44-50, January/February 2004.

47. 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.

48. 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.

49. 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.

50. 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.

51. 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.

52. 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.

53. See-Kiong Ng, Soon-Heng Tan**. Discovering Protein-Protein Interactions.** *Journal for Bioinformatics and Computational Biology,* 1(4):711-741, 2004.

54. 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.

55. Wei Chen , W.-K. Sung, **On Half Gapped Seeds**. *Proceedings of 14th International Conference on Genome Informatics*, Yokohama, December 2003.

56. 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.

57. 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.

58. 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. 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.

2. Limsoon Wong**. Exciting Media.** *Invited talk at World Scientific*, 6 April 2004.

3. Limsoon Wong. **A Practical Introduction to Bioinformatics**. *Invited short course at National Yang Ming University*, Taipei, Taiwan, 22 May 2004.

4. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments**. *Invited talk at National Yang Ming University*, Taipei, Taiwan, 26 May 2004.

5. Limsoon Wong. **Convexity of Itemset Spaces**. *Invited talk at Academia Sinica*, Taipei, Taiwan, 28 May 2004.

6. 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.

7. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments**. *Invited talk at Genome Institute of Singapore*, 4 June 2004.

8. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments.** *Invited talk at 3rd International Conference on Bioinformatics*, Auckland, New Zealand, 4-8 September 2004.

9. 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.

10. 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.

11. Limsoon Wong. **Gene Finding and Gene Feature Recognition by Computational Analysis**. *Invited tutorial at Beijing Normal University*, Beijing, 21-25 November 2004.

12. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments.** *Invited talk at Tsing Hua University*, Beijing, 25 November 2004.

13. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments.** *Invited talk at Edinburgh University*, Edinburgh, 1 October 2004.

14. 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.

15. Limsoon Wong. **Knowledge Discovery in Biomedicine**. *Invited talk at National Healthcare Group Annual Scientific Congress*, Raffles Convention Centre, Singapore, October 2004.

16. 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.

17. Limsoon Wong. **Selection of Patient Samples and Genes for Disease Prognosis.** *Invited talk at "Informatics Inspired Biology" Symposium*, BioPolis, Singapore, 16 January 2005.

18. 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.

19. 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)

Cumulative:

20. Limsoon Wong. **Assessing Reliability of Protein-Protein Interaction Experiments.** *Invited talk at Lilly Systems Biology Symposium*, BioPolis, Singapore, 4 February 2004.

21. 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.

22. 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

MILESTONES AND STATUS

A. Data mining technologies **(ongoing)**
B. Gene feature recognition **(ongoing)**
C. Gene expression analysis
D. Venome informatics
E. Pathway informatics **(ongoing)**
F. Intelligent Data Warehousing with application to Bioinformatics **(ongoing)**

Principal Investigator:

Name: ___Wong Limsoon_____

Signature: ____ _____ Date: ___25 January 2005_____

Section B: **To be completed by I²R Reviewing Officer:**

Comments: _____
_____
_____
_____
_____
_____
_____

Name: _____

Signature: _____ Date: _____