SIX-MONTHLY PROGRESS REPORT
FOR I²R-SOC JOINT LAB

Section A: To be completed by Principal Investigator

<table>
<thead>
<tr>
<th>Reporting Period</th>
<th>1 February 2005 – 31 October 2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Vote No. at SOC</td>
<td>R-252-000-172-593</td>
</tr>
<tr>
<td>Title of Project</td>
<td>Knowledge Discovery from Biological &amp; Clinical Data</td>
</tr>
<tr>
<td>PIs</td>
<td>A/P Wynne Hsu, SOC</td>
</tr>
<tr>
<td></td>
<td>Prof. Wong Limsoon, SOC/I²R</td>
</tr>
<tr>
<td>Co-PIs</td>
<td>Dr. Lee Mong Li, SOC</td>
</tr>
<tr>
<td></td>
<td>Dr. Ken Sung, SOC</td>
</tr>
<tr>
<td></td>
<td>Prof. Vladimir Bajic, I²R</td>
</tr>
<tr>
<td></td>
<td>Prof. Vladimir Brusic, I²R (Left for Univ of Queensland, Aug 05)</td>
</tr>
<tr>
<td></td>
<td>A/Prof. Ng See Kiong, I²R</td>
</tr>
<tr>
<td></td>
<td>A/Prof. Li Jinyan, I²R</td>
</tr>
<tr>
<td>Project Duration (Start Date – End Date)</td>
<td>1 July 2003 to 30 June 2006</td>
</tr>
</tbody>
</table>

1. Expenditure Level – Utilisation Rate

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

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**

<table>
<thead>
<tr>
<th>Manpower Category</th>
<th>Planned Full-time</th>
<th>Actual Full-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhD student</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>3</td>
</tr>
</tbody>
</table>

**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)**

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
• I2R funded
• Passed qualifying 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
• I2R 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
• 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 adequate 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
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”.*

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

<table>
<thead>
<tr>
<th>Capability Building Indicators</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Carrying out world class research</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of patents filed (please provide listing)</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>No. of patents granted (please provide listing)</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>No. of papers published in prestigious journals or conferences (please provide listing)</td>
<td>23 (includes: 1 book and 5 book chaps)</td>
<td>81 (include: 6 books, 14 book chaps)</td>
<td></td>
</tr>
<tr>
<td>No. of joint programmes with prestigious international research organisations</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Developing manpower</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of undergraduate/polytechnic students attached to the project for more than 3 months</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>No. of postgraduate students attached to the projects for more than 3 months</td>
<td>0</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>No. of RSEs trained through formal post-graduate studies</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>No. of RSEs trained through joint projects</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>No. of conferences &amp; seminars</td>
<td>1 (4th joint)</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>
(b) Additional milestones/deliverables not captured in 4(a) above (use additional pages if necessary)

Conferences & seminars organized  
New:

Cumulative:
2. 17-21/1/05. 3rd Asia-Pacific Bioinformatics Conference, held at I2R, 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 I2R, 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:
5. E. Huang, L. Yang, R. Chowdhary, A. Kassim, V.B. Bajic, An algorithm for ab initio


Cumulative:


36. Soon Heng Tan, See-Kiong Ng. **Discovering Protein-Protein Interactions.** *The Practical


| 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. |
| Huiqing Liu, Hao Han, Jinyan Li, Limsoon Wong. | An in silico method for prediction of...


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.


Invited talks

New:

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.


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:


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 (I2R/SOC)
Signature: ______________________ Date: 15 October 2005

Section B: To be completed by I2R Reviewing Officer:

Comments:

Name:
Signature: ______________________ Date: ________________