



Agency for  
Science, Technology  
and Research

A\*STAR  
(Agency for Science, Technology and Research)  
10 Science Park Road, #01-01/03  
The Alpha, Singapore Science Park II  
Singapore 117684  
T +65 6779 7066 F +65 6777 1711  
www.a-star.edu.sg

## **JOINT SINGAPORE-CANADA WORKSHOP ON “THE INTERFACE OF BIOLOGY WITH INFORMATION TECHNOLOGY”**

You are cordially invited to participate at the 2nd A\*STAR-NRC Workshop (17<sup>th</sup> –18<sup>th</sup> March 2003). The theme of this workshop is "Interface of Biology with Information Technology". This workshop will feature a number of invited speakers from Singapore and Canada. Attendance is free.

You are also welcomed to put up poster presentation (limit to 2 per organisation) at the workshop itself.

This workshop is part of an on-going effort to promote interaction between Singapore and Canada researchers.

### **OBJECTIVE**

To explore research areas that are likely to result in joint Singapore-Canada collaborations in the emerging field of post-genome bioinformatics. The target audience is professionals working in life sciences and bioinformatics, either as researchers, developers, or service providers.

### **DATE**

17<sup>th</sup> –18<sup>th</sup> March 2003 (2 full-day programme)

### **VENUE**

Institute for Infocomm Research (I<sup>2</sup>R)  
21 Heng Mui Keng Terrace  
Singapore 119613

### **SPONSORS**

Agency for Science, Technology and Research (A\*STAR), Singapore, and  
National Research Council (NRC), Canada

### **ORGANIZERS**

Institute for Infocomm Research (I<sup>2</sup>R), and  
BioInformatics Institute (BII)

## **REGISTRATION**

Please email (your name, organisation/department and contact telephone) to <tan\_see\_hwee@a-star.edu.sg> to confirm your participation so that a seating is reserved for you.

Please also indicate if your organisation will be interested to put up any poster presentation.

The cut-off date for registration is March 3, 2003.

## **ENQUIRIES**

If you have any questions about the workshop, please email to <tan\_see\_hwee@a-star.edu.sg>.

## Lists of Invited Speakers

### Singapore

	<b>Topic</b>	<b>Research Organisation</b>	<b>Investigators</b>
1	The SVM Classification of Light Regulated Arabidopsis Genome Expression Profiles	Nanyang Technological University	Zonglin Zhou (MPE), Jinming Li (BS), Kim Meow Liew (MPE), Chee Keong Kwoh (CE), Tet Fatt Chia (NIE)
2	Development of a system to identify and analyze protein-protein interactions	Genome Institute of Singapore	P. R. Kolatkar, Jer-Ming Chia, Kuang Yuyu
3	Comparative Genomics of Virulent and Avirulent Burkholderia Species Using Whole Genome DNA Microarrays	National Cancer Centre	Catherine Ong, Ooi Chia Huey, Dongling Wang, Hweeling Chong, May Ann Lee, and Patrick Tan
4	Grid Computing and Bioinformatics	Bioinformatics Institute	Arun Krishnan
5	Integrating Genomic and Phenotypic Information to Support Clinical Decision Making	Medical Computing Laboratory, School of Computing, National University of Singapore	Leong Tze Yun et. al.
6	Workflow integration for high throughput management of genome annotation and structural biology database curation	Bioinformatics Centre, National University of Singapore	Tan Tin Wee, S Subbiah and Shoba Ranganathan
7	Bioinformatics analysis pipeline: from genome to function	Institute for Infocomm Research, Singapore	Vladimir B. Bajic, Li Jinyan, See Kiong Ng, Vladimir Brusic

### Canada

	<b>Topic</b>	<b>Research Organisation</b>	<b>Investigators</b>
1	Class Prediction and Discovery Using Gene Microarray and Proteomics Mass Spectroscopy Data: Curses, Caveats, Cautions	Institute for Biodiagnostics (Winnipeg)	R. L. Somorjai, B. Dolenko, and R. Baumgartner
2	BioMiner – an integrated data mining infrastructure for genomics and proteomics	Institute for Information Technology (Ottawa)	Fazel Famili and Youlian Pan
3	Chemical Genomics	The Steacie Institute for Molecular Sciences (Ottawa)	John Pezacki
4	Direct Examination of Campylobacter Lipooligosaccharides, Capsules and N-linked Glycans from Whole Cells by Mass Spectrometry and High-Resolution Magic Angle Spinning NMR Spectroscopy	Institute for Biological Sciences (Ottawa)	Christine M. Szymanski, Frank St. Michael, Harold C. Jarrell, Jianjun Li, Michel Gilbert, Suzon Larocque and <u>Jean-Robert Brisson</u>
5	Identification of early cancer genes by mining mouse breast cancer microarray data	Biotechnology Research Institute (Montreal)	Edwin Wang, Herve Hogues, and Enrico O. Purisima

6	Computational tools for profiling protein-ligand binding affinities and specificities	Biotechnology Research Institute (Montreal)	Enrico O. Purisima, Mohammed Naïm, Kathryn Rankin and Traian Sulea
7	Grid technology applications for bioinformatics	Canadian Bioinformatics Resource, Institute for Marine Biosciences	Simon Mercer
8	Development of algorithms for bacterial comparative genomics using microarrays	Institute for Biological Sciences (Ottawa)	Eduardo Taboada, Catherine Carrillo and John Nash
9	Mining of genome-wide transcription profiling data to understand neurogenesis and neurodegeneration in alzheimer brains	Institute for Biological Sciences (Ottawa)	Qing Y. Liu and P. Roy Walker

Singapore

Abstracts

## **Singapore Abstract 1**

**Authors:** Zonglin Zhou (School of MPE, NTU), Jinming Li (School of BS, NTU), Kim Meow Liew (School of MPE, NTU), Chee Keong Kwoh (School of CE, NTU), Tet Fatt Chia (NIE, NTU)

**Institute:** Nanyang Technological University

**Title:** The SVM Classification of Light Regulated Arabidopsis Genome Expression Profiles

### **Description:**

The emerging technology of gene expression analysis is a promising tool for the distinction of functional classes of genes. We report for the first time that light regulated Arabidopsis gene expression profile can serve as a measure to distinguish the functional classes of Arabidopsis genes using the trained SVMs. Motivated by the good performance of binary SVM classifiers in the classification of Arabidopsis gene functions, we adopt one commonly used combination scheme: one-against-one, which combines binary SVM classifiers and constructs a bottom-up binary tree for multiclass prediction. We achieve 63.3% prediction accuracy on test samples. The prediction accuracy is limited by some factors, such as the number of samples we have, and the inherent variability of the microarray experiment.

## **Singapore Abstract 2**

**Authors:** P. R. Kolatkar, Jer-Ming Chia, Kuang Yuyu

**Institute:** Genome Institute of Singapore

**Website:** <http://www.genomeinstitute.org/homepage/gisresearchprojectsfocus.jsp>

**Title:** Development of a system to identify and analyze protein-protein interactions

### **Description:**

#### Protein-Protein Interactions Database (PPDB)

Platforms such as sequencing and microarrays can generate information about specific sequences and expression levels respectively but say nothing about the interplay between the pieces. Yeast two hybrid systems and high throughput mass spec systems have yielded large amounts of prokaryotic and yeast data (Tong et al). In this mass of information, false positives or low coverage are problems and thus a system to help augment the quality of this information will better support the functional utility of this data. The recent review from von Mering et al describes some of the different methods for protein-protein interaction analysis and the advantages and drawbacks of each.

My lab has been interested in data integration of protein-protein interactions and visualization of pathways for several years. The PPDB concept was hatched in my lab about two years ago. An initial web-enabled version has already been implemented. We have implemented a few of the features we eventually want to see in the complete PPDB in the current version. The first feature is a variation of Marcotte's Rosetta idea but we actually use domains themselves for comparisons. The use of domains rather than simply sequence similarity allows inclusion of more true positives but also more false positives. However, we next apply literature mining tools developed in the lab of Limsoon Wong at LIT(Singapore) to look for instances of the interaction pair for validation. Using this simple initial implementation, we obtained approximately 300,000 putative interactions using many species from E. coli to humans. We are currently in the process of adding structural information. We are collaborating with Peter Kuhn (Scripps) and Lin Kui (Beijing Normal University) to implement various features of this strategy.

### **Singapore Abstract 3**

**Authors:** Catherine Ong, Ooi Chia Huey, Dongling Wang, Hweeling Chong, May Ann Lee, and Patrick Tan

**Institute:** National Cancer Centre

**Title:** Comparative Genomics of Virulent and Avirulent *Burkholderia* Species Using Whole Genome DNA Microarrays

#### **Description:**

The bacterial species *Burkholderia pseudomallei* and *B. mallei* are the causative agents of the human diseases melioidosis and glanders respectively, and both species are regarded by military authorities as potential biowarfare agents. We used whole genome *B. pseudomallei* DNA microarrays containing approximately 6900 predicted genes to compare the genomic content of several *B. pseudomallei* and *B. mallei* strains, as well as the related but clinically avirulent species *B. thailandensis*. At a 99.99% confidence level, we identified 376 and 325 predicted genes that were deleted in *B. mallei* and *B. thailandensis* respectively compared to the *B. pseudomallei* reference strain K29643. Secondary validation of a subset of these array results was achieved either by PCR or by comparison to previously published reports. Tiling of the array probes across the *B. pseudomallei* genome revealed that many of the *B. mallei* deletions occur in the form of large contiguous DNA stretches, particularly on Chromosome 2. In contrast, deletions observed in *B. thailandensis* were randomly dispersed throughout the *B. pseudomallei* genome, suggesting a distinct mechanism of divergence compared to *B. mallei*. Previously unidentified genes that were deleted between the three bacterial species were associated with a wide variety of cellular functions, such as nitrogen fixation, polyketide formation, and amino acid biosynthesis. Our results raise several hypotheses regarding the molecular mechanisms underlying the diverse environmental and biological properties exhibited by members of the *Burkholderia* genus, and constitute a robust framework upon which to interpret future functional genomic and proteomic information regarding these species.

## **Singapore Abstract 4**

**Authors:** Arun Krishnan

**Institute:** BioInformatics Institute

**Website:** <http://www.bii.a-star.edu.sg/> arun

**Title:** Grid Computing and Bioinformatics

### **Description:**

Improvements in the performance of processors and networks have made it feasible to treat collections of workstations, servers, clusters and supercomputers as integrated computing resources or Grids. However, the very heterogeneity that is the strength of computational and data grids can also make application development for such an environment extremely difficult. The talk will introduce the concept of grid computing as well as present a grid enabled, high-throughput version of a bioinformatics application, BLAST. BLAST is a sequence alignment and search technique that is embarrassingly parallel in nature and thus amenable to adaptation to a grid environment. The application has been tested on a “mini-grid” testbed and the results presented here show that for large problem sizes, a distributed, grid-enabled version can help in significantly reducing execution times. The talk will also discuss other grid-related projects that are being carried at BII.

## **Singapore Abstract 5**

**Authors:** Leong Tze Yun et. al.

**Institute:** Medical Computing Laboratory, School of Computing, National University of Singapore

**Title:** Integrating Genomic and Phenotypic Information to Support Clinical Decision Making

### **Description:**

We report on the recent initiatives and on-going investigations at the Medical Computing Laboratory to combine genomic and phenotypic information to support clinical decision making. Some of the on-going projects include:

Gene-profiling for customized therapy:

The objective is to understand whether gene profiling is helpful in detecting cancer patients who will get good response to chemo-radiotherapy. The “marker” genes which display the difference in expression level for different patients will be identified. These markers can be used to predict the patients’ responses to therapy. Such gene profiling information, together with patient history and other clinical records, will be incorporated into clinical decision models for customized therapy planning.

Knowledge discovery in genome sequences:

Allelic association occurs due to linkage disequilibrium where alleles that are very close to the disease susceptibility allele are inherited together over many generations, such that the same set of alleles are detected in many affected individuals which are not apparently related. The objective is to discover the locations of disease genes from observed associations between marker alleles and disease phenotypes. Accurate identification of disease gene location would lead to more accurate diagnosis and prognosis of the disease. Genomics and proteomics studies routinely depend on similar (homology) searches based on the strategy of finding short sequence matches (seed) which are then extended. Increasing seed size decreases sensitivity whereas decreasing seed size slows down computation. The objective is to examine the seed design that could offer good efficiency performance in the search while maintaining high sensitivity in retrieving the related sequences that are relevant given a query sequence.

Translation initiation site (TIS) is the position in cDNA sequence to start constructing proteins. Accurate recognition of TIS can help us understand the gene structure and its function. The objective is to identify the translation initiation site in each cDNA sequence to support gene structure recognition and functional derivation.

## **Singapore Abstract 6**

**Authors:** Tan Tin Wee, S Subbiah and Shoba Ranganathan

**Institute:** National University of Singapore

**Website:** <http://www.bic.nus.edu.sg/>

**Title:** Workflow integration for high throughput management of genome annotation and structural biology database curation

### **Description:**

Since 1996, our researchers have been working on various aspects of database integration which had led to commercialisation of broadscale database integration with pathway visualisation tools.

However, from 1998, we have worked on a step beyond, which is to look at how the entire workflow process of biological research which involves interfacing and information exchange amongst researcher, device and database. This has led to an enterprise application integration initiative which has a workflow integration system in which each functional node can be represented as an enterprise-wide browsable object, which can be invoked in a Java workflow diagramming GUI, and executed sequentially or in a scheduled manner to run the processes in the workflow nodes on different remote machines. We have attempted to deploy this workflow in the context of an campus intranet NUS BioGrid and a wide area Asia Pacific BioGrid.

A set of more than 200 bubbles, including 160 from the EMBOSS suite, 40 from PHYLIP, 20 Unix utilities, a dozen from MySQL, a few for Globus Grid, BLAST, FASTA and CLUSTALW have been constructed. See <http://www.apbionet.org/pr/cray-lion-koopprime.html> and <http://www.apbionet.org/pr/apbionet-cray-lion-koopprimeJul02news.html>.

We are in the process of :

- a. deploying this workflow in the context of a more complex package of software in the context of the Asia Pacific Bioinformatics Network's APBioGrid project,  
See <http://www.apbionet.org/pr/apbionet-idrc-grant.html>
- b. targeting specific biological problems to prove the utility and applicability of such semi-automated workflows for handling a specific genome annotation project and structural biology databases.

The genome project targetted is the annotation of the already sequenced genome of *Burkholderia pseudomallei*. This is a Gram negative bacterial

pathogen which is found in the Southeast Asian and northern Australian soil, and represents a potential bioterrorism threat because of its high resistance against antibiotics, its highly dangerous fulminant form in causing septicaemia, and its ability for latency and recrudescence.

For sometime, biologists at NUS have studied this organisation at the Molecular level, including molecular typing, strain differentiation, gene cloning of virulence factors, immunological responses to infection and a *C.elegans* infection model and in collaboration with others, microarray analysis of expression profiles of the bacterium. Recently as the genome has been entirely sequenced at the Sanger Centre and the data published on the Web, we have been able to embark on genome annotation of this genome. As genome annotation involves the interplay of researcher with many different databases in a workflow dependent manner, this represents a good test case with possibility of validation in the wet laboratory context in combination with Grid Computing, (See <http://www.bic.nus.edu.sg/biogrid/list.html>) and hopefully to interface with the microarray group doing the expression analysis.

At the same time, we have been working on intron-exon databases which are cross-related to structure. This involves very complex interaction between researcher and database integration, as well as molecular modelling techniques. Therefore, to speed up and to make the process more efficient, we have been trying to explore ways of streamlining the workflow of database curation using the workflow integration system described above. See <http://surya.bic.nus.edu.sg/xdom/>

Work is in progress to discover more test cases where we can further optimise the workflow integration platform, and refine its Java and RMI interfaces to use XML, SOAP and UDDI, and to inter-operate with latest versions of Globus Toolkit.

## **Singapore Abstract 7**

**Authors:** Vladimir B. Bajic, Li Jinyan, See Kiong Ng, Vladimir Brusic

**Institute:** Institute for Infocomm Research, Singapore

**Title:** Bioinformatics analysis pipeline: from genome to function

### **Description**

The I2R bioinformatics projects which are logically linked focus on the three main fields: analysis of functional sites in DNA, prediction of gene expression, and protein functional analysis.

For prediction of functional sites in DNA we have developed Dragon suite of programs. The general underlying technology consists of systems of sensors that capture relevant signals from the raw DNA, process these using artificial neural networks and other classification tools, and predict functional sites. The high accuracy predictions were achieved in the prediction of promoters, transcription initiation sites, and poly-A sites.

For classifying gene expression profiles or other types of medical data for medical diagnostics, simple rules are preferable to non-linear distance or kernel functions. This is because rules may help researchers and medical doctors to understand more about the application in addition to accurate classification. We have devised a highly accurate and understandable classifier, named PCL (Prediction by Collective Likelihood) based on the concept of emerging patterns. Using PCL, we have successfully discovered novel rules that describe the gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL), as well as proteomic mass/charge profiling data of ovarian cancer.

The prediction of protein function is based on reclassification of structural groups of proteins based on observed functional properties, application of sequence similarity searches, and the definition of rules for assigning class membership of query sequences. The whole process is facilitated using BioWare – a platform for biological data warehousing. Using this approach we have built datanbases of scorpion, marine snail, and snake toxins which support functional annotation of new sequences.

# Canada

## Abstracts

## **Canada Abstract 1:**

**Authors:** R. L. Somorjai, B. Dolenko, and R. Baumgartner

**Institute:** Institute for Biodiagnostics (Winnipeg)

**Website:** <http://www.ibd.nrc-cnrc.gc.ca/>

**Title:** Class Prediction and Discovery Using Gene Microarray and Proteomics Mass Spectroscopy Data: Curses, Caveats, Cautions

### **Description:**

Two practical realities constrain the analysis of microarray data, mass spectra from proteomics, and biomedical infrared or magnetic resonance spectra. One is the “curse of dimensionality”: the number of features characterizing these data is in the thousands or tens of thousands. The other is the “curse of dataset sparsity”: the number of samples available for analysis is limited. The consequences of these two curses are far-reaching when such data are used to classify the presence or absence of disease.

Using very simple classifiers, we show for several publicly available microarray and proteomics datasets how these curses influence classification outcomes. In particular, even if the sample per feature ratio (SFR) can be increased by feature extraction/reduction methods to the recommended 5-10, dataset sparsity can render any classification result statistically suspect. A further important consequence is that several sets of “optimal” features are typically identifiable for sparse datasets, all producing classifiers with zero misclassification errors, both for the training and independent validation sets. This non-uniqueness of the feature sets leads to interpretational difficulties and casts doubt on the validity of assertions regarding the biological relevance of any specific “optimal” feature set, when deduced from a sparse dataset. We suggest a possible approach to assess the relative quality of classifiers that are apparently equally good.

## **Canada Abstract 2:**

**Authors:** A. Fazel Famili and Youlian Pan

**Institute:** Institute for Information Technology (Ottawa)

**Website:** <http://iit-iti.nrc-cnrc.gc.ca>

**Title:** BioMiner – an integrated data mining infrastructure for genomics and proteomics

### **Description:**

With the advancements in genomics and proteomics, enormous amount of data have been generated. This has created a demand for integrated data mining systems. BioMiner, a response to such a demand, has resulted from collaboration between IIT (Institute for Information Technology) and IBS (Institute for Biological Sciences) at NRC. This is a modular, multi-layer system, and designed as a suite of data mining tools for both research and development. The available functionalities include, but are not limited to, data quality checking, data characteristics checking (both for discovering hidden anomalies), data filtering, 2-D and 3-D data visualization including virtual reality, data re-representation, clustering with 28 algorithms, several pattern recognition algorithms, and sequence motif finding. Other modules including automatic retrieval of external information (e.g. metabolic networks, gene regulation networks), and comparison of genes in the data set, using local and external databases will soon be developed. We are in the process of incorporating parallel high performance programming into BioMiner. The existing architecture is designed to accommodate more modules in response to the development of research in genomics, proteomics and data mining technology.

Colleagues at IBS and PBI have used BioMiner to discover new knowledge from their research data, and presented at the GHI-2002 in Ottawa. External collaboration with Children Hospital of Eastern Ontario is in progress. We have applied BioMiner in analyzing published leukemia data (Golub et al, Science 286: 531-537, 1999) and discovered new features that were not reported by the original authors. This result was presented at GHI-2002 and will be presented at the International Conference in Applied Informatics (Feb. 2003) in Austria. This presentation will focus first on our lessons learned from this collaborative project. We will then provide a few examples about our data mining case studies. With experience in the development of BioMiner and our expertise in data mining and functional genomics, we will certainly be interested in defining new collaborative research directions within the BioMine project. For more information please refer to the BioMine project home page ([http://iit-iti.nrc-cnrc.gc.ca/biomine\\_e.trx](http://iit-iti.nrc-cnrc.gc.ca/biomine_e.trx)) and contact the project leader Dr. Fazel Famili ([fazel.famili@nrc-cnrc.gc.ca](mailto:fazel.famili@nrc-cnrc.gc.ca)).

### **Canada Abstract 3:**

**Authors:** John Pezacki

**Institute:** The Steacie Institute for Molecular Sciences (Ottawa)

**Website:** <http://steacie.nrc-cnrc.gc.ca/>

**Title:** Chemical Genomics

#### **Description:**

My work in the areas of chemical biology and functional genomics rely heavily on new and creative methods for data analysis, information technology, and systems biology. I would therefore be very excited to participate in the Singapore workshop and to have the opportunity to develop collaborations. I have expertise in the area of chemical approaches to disruption of gene transcription through interference with transcription factor/DNA interactions; the fundamental processes that control gene activation (Chem. and Biol. 2002, 9, 821). I am also interested in designing new types of small molecules that can act as functional genomics tools and powerful probes into biological processes. Some of my group's current work involves designing probes for investigating the host-virus interactions of HCV infections (for a representative publication see: Proc. Natl. Acad. Sci USA, 99, 15660). I would contribute to the workshop by presenting some recent work from my group and discussing potential areas of collaboration in the areas of chemical genomics. My interests lie specifically in developing new bioinformatics approaches to discovering mechanisms of action of small molecules (e.g. natural products, metabolites, synthetic materials) on the complex signaling pathways within living cells.

#### **Canada Abstract 4:**

**Authors:** Christine M. Szymanski, Frank St. Michael, Harold C. Jarrell, Jianjun Li, Michel Gilbert, Suzon Larocque and Jean-Robert Brisson

**Institute:** Institute for Biological Sciences (Ottawa)

**Website:** <http://ibs-isb.nrc-cnrc.gc.ca/>

**Title:** Direct Examination of *Campylobacter* Lipooligosaccharides, Capsules and N-linked Glycans from Whole Cells by Mass Spectrometry and High-Resolution Magic Angle Spinning NMR Spectroscopy

#### **Description:**

Glycomics, the study of microbial polysaccharides and genes responsible for their formation, has become an increasingly important area of investigation. Methods for the direct analysis of bacterial sugars from whole cells of various *Campylobacter* serostrains and mutants are outlined. Using capillary-electrophoresis coupled with sensitive electrospray mass spectrometry, we demonstrate extensive variability in the lipid A component of *C. jejuni* lipooligosaccharides. In addition, several differences in LOS core structures that were not observed previously are also described. High-resolution magic angle spinning (HR-MAS) NMR was used to examine capsular polysaccharides from *Campylobacter* whole cells and showed profiles similar to that observed with purified polysaccharides analysed by solution NMR. This method also exhibited the potential for *Campylobacter* serotyping, mutant verification, and preliminary sugar analysis. Examination of growth from individual colonies of *C. jejuni* NCTC 11168 by HR-MAS NMR demonstrated that the capsular - glycan modifications are phase variable and result in differential silver-staining patterns and reactivity with immune sera on DOC-PAGE. Interestingly, this method also detected the N-linked glycan, GalNAc- $\alpha$ 1,4-GalNAc- $\alpha$ 1,4-[Glc $\alpha$ 1,3-]GalNAc- $\alpha$ 1,4-GalNAc- $\alpha$ 1,4-GalNAc- $\alpha$ 1,3-Bac- $\alpha$ 1,N-Asn-Xaa, where Bac is bacillosamine, 2,4-diacetamido-2,4,6-trideoxy-D-glucopyranose which was present in all *C. jejuni* and *C. coli* isolates analysed. This is the first report of HR-MAS NMR detection of N-linked glycans on glycoproteins from intact bacterial cells. Protein N-glycosylation was abolished when the *pglB* gene was mutated, providing further evidence that the enzyme encoded by this gene is responsible for formation of the glycopeptide N-linkage. Comparison of the *pgl* locus with that of *Neisseria meningitidis* suggested that most of the homologous genes are probably involved in the biosynthesis of bacillosamine.

## **Canada Abstract 5:**

**Authors:** Edwin Wang, Herve Hogues, and Enrico O. Purisima

**Institute:** Biotechnology Research Institute (Montreal)

**Website:** [http://www.bri.nrc-cnrc.gc.ca/rd/index\\_e.html](http://www.bri.nrc-cnrc.gc.ca/rd/index_e.html)

**Title:** Identification of early cancer genes by mining mouse breast cancer microarray data

### **Description:**

A mouse mammary epithelial cell line, BRI-JM01, when treated with TGF- $\beta$ , will initiate and develop breast cancer. The transcriptome changes involved in the cancer progression induced by TGF- $\beta$  in BRI-JM01 cells over a time-course of 2, 4, 6, 12, 24 hrs were analyzed. Using microarray containing 15,264 verified mouse ESTs, 284 genes whose expression levels were significantly modulated were identified.

In order to find early genes that trigger or regulate the cancer progression, we constructed a Markov chain-based gene network that may reflect the casual relationships among these significantly modulated genes. The gene network was analyzed to identify different gene groups. To explore the knowledge domain of these gene groups in the gene network, a literature network was constructed based on the relationships among these genes and related genes in literature. Genes in the gene network were further annotated using Gene Ontology (GO), Locuslink, and GeneCard. Eighteen potential early genes were identified, half of these genes are function-inferred and are transcription factors, and others are ESTs with unknown function. This is in agreement with the hypothesis that most early cancer genes encode transcription factors.

We also identified more than 20 tumor suppressor related genes using the nearest-neighbor technique and Bayesian-Markov chain-based dynamic clustering technique and using known tumor suppressor genes as a query gene set. Comparison with the National Cancer Institute (NCI) SAGE human breast cancer data validated some of these suppressor related genes. The identified genes in this study will be subjected to further experimental validation in the laboratory.

## **Canada Abstract 6:**

**Authors:** Enrico O. Purisima, Mohammed Naïm, Kathryn Rankin and Traian Sulea

**Institute:** Biotechnology Research Institute (Montreal)

**Website:** [http://www.bri.nrc-cnrc.gc.ca/rd/index\\_e.html](http://www.bri.nrc-cnrc.gc.ca/rd/index_e.html)

**Title:** Computational tools for profiling protein- ligand binding affinities and specificities

### **Description:**

Protein-protein and protein- ligand interactions are at the core of the molecular recognition process. The ability to dissect out the thermodynamic determinants of protein- ligand binding is extremely valuable in elucidating the molecular mechanisms at play and in providing strategies for modifying or modulating these interactions. We have developed and applied a variety of computational tools for analyzing and predicting the potency and specificity of binding interactions. These techniques rely on a detailed calculation of interaction energies coupled with a high quality solvation model. We will present our current work in three areas: 1) development of scoring functions for virtual screening of compound libraries, 2) visualization of charge complementarity profiles on protein surfaces, 3) in silico mutagenesis studies (e.g., virtual alanine scanning) of the contribution of individual amino acids to binding affinity.

## **Canada Abstract 7:**

**Authors:** Simon Mercer

**Institute:** Canadian Bioinformatics Resource, Institute for Marine Biosciences

**Websites:** <http://cbr-rbc.nrc-cnrc.gc.ca> (CBR)  
<http://www.imb.nrc.ca/> (IMB)

**Title:** Grid technology applications for bioinformatics

### **Description:**

Grid technology offers the potential for the establishment of a uniquely scaleable infrastructure, which in turn represents one of the few credible strategies to address the explosion of biological data requiring storage, analysis and dissemination in the postgenomic era. In addition, this emerging technology leverages existing hardware investments, prolongs the effective life of existing compute resources, and promotes entirely new modes of collaboration between researchers worldwide.

The experimental nature of existing grid software (for example Avaki and the Globus toolkit) leaves much to be desired with respect to stability, security and ease of implementation - all prerequisites for wide-scale academic or commercial deployment. The Canadian Bioinformatics Resource therefore proposes that, in conjunction with the Singapore Bioinformatics Institute, we partner with Invio, a bioinformatics startup company located in Halifax, NS for the production of a wide-scale user management tool to assist in the deployment of a grid, with the intention that the resultant software package be made freely available as part of the Globus grid toolkit, the pre-eminent public domain grid toolkit.

This project would confirm the commitments of both Canada and Singapore to grid developments, and ensure both retained their place at the forefront of practical grid research and deployment. Contributions to Globus will (literally) set the standard for grid architecture for years to come.

## **Canada Abstract 8:**

**Authors:** Eduardo Taboada, Catherine Carrillo and John Nash

**Institute:** Institute for Biological Sciences (Ottawa)

**Website:** <http://ibs-isb.nrc-cnrc.gc.ca/>

**Title:** Development of algorithms for bacterial comparative genomics using microarrays

### **Description:**

The use of microarrays for comparative genomics (“genomotyping”) offers benefits over current methodologies, (e.g. serotyping, RFLPs) which rely on arbitrary phenotypes to infer genetic relationships. In genomotyping, gene conservation is inferred from signal intensity. Current array-analysis algorithms are biased towards interpreting data from expression profiling experiments where signal intensity is a measure of transcript abundance. We propose to develop algorithms that incorporate empirical data based on the relationship between hybridization stringencies, sequence divergence and signal strength. Genomotyping data is of an inherently evolutionary nature and current clustering algorithms, which fail to incorporate an evolutionary model, limit our ability to properly infer phylogenetic relationships based on whole genome array data. Thus we also propose to develop clustering methodology to overcome these limitations. These proposed approaches would further develop the applicability of array-based methods in comparative genomics.

Functional genomics of prokaryotes have different emphases from those of eukaryotes. As part of our studies on the virulence of *Campylobacter jejuni*, we have produced a whole-genome DNA microarray to study comparative genomics and lateral gene transfer in this pathogen. In addition to our development work in genomotyping, our group is developing a data pipeline integrating gene annotation with microarray results from genomotyping and expression profiling of *C. jejuni* from host-pathogen studies, and eventually with proteomic profiling, including metabolomic studies.

## **Canada Abstract 9:**

**Authors:** Qing Y. Liu and P. Roy Walker

**Institute:** Institute for Biological Sciences (Ottawa)

**Website:** [<http://ibs-isb.nrc-cnrc.gc.ca/>](http://ibs-isb.nrc-cnrc.gc.ca/)

**Title:** Mining of genome-wide transcription profiling data to understand neurogenesis and neurodegeneration in alzheimer brains

### **Description:**

Genome wide transcription profiling is a powerful technique for studying the enormous complexity of cellular states. To better understand the neuronal phenotype we have been profiling changes in gene expression during neurogenesis in Human Embryonal Carcinoma cells using Human NT2 EC cells induced to differentiate by the morphogen, retinoic acid. Interpretation of the data generated by genome-wide transcription profiling has proven to be particularly challenging. To overcome this we have developed Biominer, a suite of data mining tools, coupled to virtual reality representations of the activity of the genome. In this way we can examine, simultaneously, the changes in the more than 3,000 genes that are differentially expressed as the EC cells are transformed into neurons. More specifically, by coupling data mining with literature mining, we are gaining a comprehensive understanding of the transcriptional regulation of the major functional processes occurring in neuronal cells.

The same approach applied to disease tissue may reveal quantitative and qualitative alterations in gene expression that give information on the context or underlying basis for the disease. Our approach addresses the problem of dealing with large amounts of gene expression data through dimension reduction and elimination of irrelevant or less promising attributes. It reveals how models that can be used in future classifications and diagnosis of diseases can be generated from relatively small gene expression data sets. These models will allow medical researchers to develop improved tools for disease classification and diagnosis.

A method complementary to DNA microarray analysis is subtractive hybridization. We have employed a novel subtractive technique to isolate unknown, rare, cell-death related genes in Alzheimer brains, followed by DNA custom microarray analysis with DNA amplicons from the subtracted cDNA libraries against a number of postmortem human brain tissues to identify genes that are uniquely associated to the disease. Using data and literature mining approaches we hope to identify not just individual genes that are altered in disease tissue, but to establish relationships between these genes that can be interpreted in terms of the genetic networks and pathways with the goal of developing a capability to modulating or replace lost functionality in impaired neurons.