Second Bioinformatics Symposium

Ottawa-Carleton Joint Collaborative Program in Bioinformatics

August 28, 2009, at the University of Ottawa

The Ottawa-Carleton Institutes of Biology, of Computer Science, of Mathematics & Statistics and the Biochemistry, Microbiology & Immunology graduate programs at the University of Ottawa are proud to invite you to a **one-day symposium** to mark the start of the academic year for the joint collaborative MSc program in Bioinformatics. This year's **plenary speaker features Dr. Hervé Philippe**, a world-renowned expert in phylogenomics at the Université de Montréal.

The idea behind this symposium is essentially to create a graduate and post-graduate community of people who work in bioinformatics and related fields, either as users or as developers of new methods. Consequently, this symposium is open to anybody and is not limited to students enrolled in the Bioinformatics program.

The symposium will be held on **Friday August 28, 2009, from 9-4:00 in SITE C0136** at the University of Ottawa (see map on last page). There will be an onsite BBQ for lunch (see map), organized by the Graduate Students' Association of the University of Ottawa.

Registration is free but mandatory. Graduate students (MSc, PhD) and PDFs with interests in Bioinformatics are strongly encouraged to register as soon as possible, and to give a presentation. Each presentations will be 30 minutes long (*including* discussion).

Inquiries and registration should be directed to Stephane Aris-Brosou (`sarisbro@uottawa.ca`). Details are also available online at `www.bioinformatics.uottawa.ca`.

1 Program

Time	Presenter	Affiliation	Title
09:00-09:10	Stéphane Aris-Brosou	BIO	Opening remarks.
09:10-10:00	Hervé Philippe	UdeM	Resolving the phylogeny and divergence time of animals: more data or better methods?
10:00-10:30	Eric Paquet	SITE	Three-dimensional indexing and retrieval from the Protein Data Bank.
10:30-11:00	**refreshments**		
11:00-11:30	Robert Davies	MATH	On the construction of a classification algorithm from genome wide association data.
11:30-12:00	Gareth Palidwor	BIO	Why do the frequencies of some G-ending codons decrease with increasing GC bias?
12:00-01:30	**lunch break**	(BBQ)	*In the courtyard between Marion and CAREG*
01:30-02:00	Manon Ragonnet	BIO	Phylogenetic analysis of population based HIV drug surveillance data may identify epidemic drivers.
02:00-02:30	Zahra Montazeri	BMI	Shrinkage estimation of expression fold change as an alternative to testing hypotheses of equivalent expression
02:30-03:00	**refreshments**		
03:00-03:30	Corey Yanofsky	OISB	Validation of differential gene expression algorithms: Application comparing fold-change estimation to hypothesis testing.
31:30-04:00	Ashkan Golshani	BIO	Towards predicting a global protein-protein interaction map of human cells
04:00-04:15	Stéphane Aris-Brosou	BIO	Closing remarks.

2 Abstracts

1. **Robert Davies**, *On the construction of a classification algorithm from genome wide association data*: Genome Wide Association (GWA) studies are large case control studies whose primary purpose is the identification and quantification of the effect that common genetic variants have on the inheritance of complex diseases. As these studies become larger and encompass more of the genome, the possibility of developing algorithms which attempt to predict whether or not a person will develop a disease based on their genome becomes more realistic. Here, we show the results of the application of a simple classification algorithm, the Nave Bayes classifier, to GWA data and discuss some of the statistical issues surrounding its implementation.

2. **Eric Paquet** & Herna Viktor , *Three-dimensional indexing and retrieval from the Protein Data Bank*: In this talk, we describe a new system for indexing and searching the three-dimensional shapes of various representations of proteins. Our approach is twofold. The first method is based on four (4) bi-dimensional views, which allows for the analysis of both the shape and the encoded physicochemical properties. The second technique is purely three-dimensional, where we explore the shape of the protein per se. Results are presented against 30.000 entries of the Protein Data Bank. We show that the system is efficient and accurate, with a retrieval time of less than a second, when performing an exhaustive search.

3. **Gareth Palidwor**, *Why do the frequencies of some G-ending codons decrease with increasing GC bias?*: Arginine and leucine codons are unusual, subject to GC changing synonymous substitutions in both the first and third codon positions. Codon usage in response to GC biased mutational pressure in these codon families should therefore differ from all others. We create a Markov chain model of GC biased synonymous substitutions for leucine and arginine codons to study their properties. The model predicts that codons with one C or G at the first or third codon position will increase in frequency with GC-biased mutation only in the low GC range, and then decrease with more GC-biased mutation. We tested this prediction using prokaryotic genomes with a wide rage of GC biases, and human genes from GC rich and poor isochores. This empirical data strongly supports the prediction of the model. The perception that C-ending and G-ending codons should increase with GC-based mutation is not generally true.

4. **Manon Ragonnet**, *Phylogenetic analysis of population based HIV drug surveillance data may identify epidemic drivers*: Drug resistance testing has generated an abundance of HIV genetic sequences that may have additional public health value. HIV pol sequences were generated from 876 serum specimens collected from drug nave, first time HIV positive patients in British Columbia between 2002 and 2005. Relationships among sequences from 2002 were inferred using neighbour-joining analysis, and clusters of infections were identified. The entire 2002-2005 dataset was then reanalysed to evaluate the relationship of subsequent infections to those identified in 2002. In 2002, Aboriginal ethnicity and intravenous drug use were correlated, and both were linked to cluster membership. Al-

though cluster growth between 2002 and 2005 was correlated with the size of the original cluster, more related infections were found in clusters seeded by the non-clustered 2002 infections. Finally, all high growth clusters were seeded from infections that were much more likely to be recent. We propose this cross-sectional analysis of existing sequences within public health databases may be useful in predicting trends in HIV transmission at the population level.

5. **Zahra Montazeri**, *Shrinkage estimation of expression fold change as an alternative to testing hypotheses of equivalent expression*: On the basis of two distinct simulation studies and data from different microarray studies, we systematically compared the performance of several estimators representing both current practice and shrinkage. We find that the threshold-based estimators usually perform worse than the maximum-likelihood estimator (MLE) and they often perform far worse as quantified by estimated mean-squared risk. By contrast, the shrinkage estimators tend to perform as well as or better than the MLE and never much worse than the MLE, as expected from what is known about shrinkage. Based on the ability of the latter to leverage information across genes, we conclude that the use of the local-FDR estimator of the fold change instead of informal or threshold-based combinations of statistical tests and non-shrinkage estimators can be expected to substantially improve the reliability of gene prioritization at very little risk of doing so less reliably.

6. **Corey Yanofsky**, *Validation of differential gene expression algorithms: Application comparing fold-change estimation to hypothesis testing*: Sustained research on the problem of determining which genes are differentially expressed on the basis of microarray data has yielded a plethora of statistical algorithms, each justified by theory, simulation, or ad hoc validation and yet differing in practical results from equally justified algorithms. Recently, a concordance method that measures agreement among gene lists have been introduced to assess various aspects of differential gene expression detection. This method has the advantage of basing its assessment solely on the results of real data analyses, but as it requires examining gene lists of given sizes, it may be unstable. We have developed two complementary methods, a cross-validation method and a posterior predictive method, that preserve the key advantage of the concordance method but do not depend on gene lists. As a nonparametric method of estimating prediction error from observed expression levels, cross validation provides an empirical approach to assessing algorithms for detecting differential gene expression that is fully justified for large numbers of biological replicates. Because it leverages the knowledge that only a small portion of genes are differentially expressed, the posterior predictive method is expected to provide more reliable estimates algorithm performance, allaying concerns about limited biological replication. In practice, the posterior predictive method can assess when its approximations are valid and when they are inaccurate. Under conditions where its approximations are valid, it corroborates the results of cross validation. Both comparison methodologies are applicable to both single-channel and dual-channel microarrays. For the data sets considered, estimating prediction error by cross validation demonstrates that empirical Bayes algorithms based on the lognormality assumption tend to outperform algorithms based

on selecting genes by their fold changes. The posterior predictive assessment confirms the poor performance of fold change selection.

7. **Ashkan Golshani**, *Towards predicting a global protein-protein interaction map of human cells*: A major goal of molecular systems biology is to uncover the protein-protein interaction (PPI) map of a cell. PPIs mediate various aspects in the structural and functional organization of a cell, including multi-faceted responses to the internal and external stimuli. Elucidating PPIs can also help us better understand the biology of complex diseases, such as cancer and diabetes, and facilitate development of certain therapeutics to alter/modify target PPIs. While large-scale experimental approaches have generated large collections of experimentally determined PPIs, a lack of an overlap between different datasets collected by the same and/or different techniques, suggests that a significant portion of cellular PPIs are not amenable to the traditional experimental procedures. In this context, computational tools hold a promise to help uncover some of these novel PPIs. Very recently, we demonstrated that PPIs in yeast cells can be successfully predicted from their primary sequences. Our approach is based on short (smaller than 25 amino acids) re-occurring polypeptide regions which are present in the dataset of interacting proteins. Here, we demonstrate that the same approach can be applied to investigate the "global" PPI map in human cells. Our current prediction sensitivity is 23% with a specificity of 99.95%. To date, from our predicted interactions for chromosome-associated proteins, we have predicted a series of novel functions for different proteins. Two of our predicted functions are for two uncharacterized human open reading frame that we termed DDRP1 and DDRP2 for DNA damage repair proteins 1 and 2. Using a plasmid-based DNA repair assay, we confirmed a role for the homolog of these proteins in yeast cells. Additionally, we have made a series of novel and interesting prediction for the HIV and Inf A virus proteins with different human proteins.