COLLOQUE BIO-INFORMATIQUE
ROBERT CEDERGREN
COLLOQUIUM IN BIOINFORMATICS

PROGRAMME

Université de Montréal
5-6 novembre 2009

Présentations orales et affiches
Poster and oral presentations

IBM

IRSC CIHR
Instituts de recherche en santé du Canada
Canadian Institutes of Health Research

CENTRE ROBERT-CEDERGREN
BIO-INFORMATIQUE et GÉNOMIQUE
UNIVERSITÉ de MONTRÉAL

Université de Montréal

Bienvenue au 6e colloque bio-informatique Robert-Cedergren !

Ce colloque, supporté par le programme de bourses d'excellence biT, une initiative stratégique des Instituts de recherche en santé du Canada (IRSC), se veut le rendez-vous annuel de la communauté universitaire oeuvrant en bio-informatique. L'objectif principal est de partager les derniers développements en ce domaine par le biais d'affiches et de présentations orales et de rendre compte de l'importance grandissante de la bio-informatique dans les sciences de la vie, incluant le domaine de la recherche en santé.

En cette sixième édition du colloque, les conférenciers invités sont :

- **Laxmi Parida,** Watson Research Center, IBM
- **John Quackenbush,** Dept. of Biostatistics, Harvard School of Public Health
- **Eran Segal**, Dept. of Computer Science and Applied Mathematics, The Weizmann Institute of Science
- **Gary Stormo**, Div. of Biology and biomedical sciences, School of Medicine Washington.

Des prix seront décernés dans les catégories suivantes :

	Meilleures présentations orales	Meilleures affiches
2e cycle	1000 $	500 $
3e cycle	1000 $	500 $
Postdoctorat	1000 $	500 $

Un excellent colloque à tous et à toutes !

Gertraud Burger, Ph.D.
Co-responsable des programmes
de 2e et 3e cycle – Bio-informatique

Welcome to the 6th annual Robert Cedergren Colloquium in Bioinformatics!

This sixth Colloquium, supported by the program biT fellowships for excellence, a strategic initiative from Candian Institutes of Health Research (CIHR), is an annual event gathering the universitary community working in Bioinformatics. The main purpose of this event is to share the latest advancement in Bioinformatics by posters and oral presentations, and to demonstrate the increasing role of Bioinformatics for life sciences in general and specifically in health research.

Keynote speakers will be:

- **Laxmi Parida,** Watson Research Center, IBM
- **John Quackenbush,** Dept. of Biostatistics, Harvard School of Public Health
- **Eran Segal**, Dept. of Computer Science and Applied Mathematics, The Weizmann Institute of Science
- **Gary Stormo**, Div. of Biology and biomedical sciences, School of Medicine Washington.

Awards will be given in the following categories:

	Best oral presentations	Best posters
M.Sc.	$ 1000	$ 500
Ph.D.	$ 1000	$ 500
Post-doc.	$ 1000	$ 500

Enjoy the Colloquium!

Gertraud Burger, Ph.D.
Leader
Bioinformatics graduate programs

Comités/Committees

Comité scientifique / Scientific Committee

Gertraud Burger, Biochimie, UdeM
Raphaël Gottardo, IRCM, UdeM
Franz Lang, Biochimie, UdeM
François Major, DIRO, UdeM
Marcel Turcotte, U.Ottawa

Arbitres / Referees

Nicolas Lartillot, Biochimie, UdeM
Vladimir Makarenkov, Informatique, UQAM
Emmet O'Brien, Biochimie, UdeM
Hervé Philippe, Biochimie, UdeM
Reza Salavati, Parasitology, McGill
Marcel Turcotte, Computer science, U.Ottawa

Modérateurs / Session Chairs

Mathieu Blanchette, Computer Science, McGill
Raphaël Gottardo, IRCM, UdeM
B. Franz Lang, Biochimie, UdeM
François Major, DIRO, UdeM

Comité organisateur / Organizing Committee

Gertraud Burger, Biochimie, UdeM
Marie Robichaud, Biochimie, UdeM

Assistance technique / Technical support

Philipe Lampron, Biochimie, UdeM
Elaine Meunier, Biochimie, UdeM

Renseignements généraux / General information

Accueil / Registration

L'accueil des participants sera situé près de la salle S1-139 du Pavillon Jean-Coutu (2940 Chemin de la polytechnique), le jeudi 5 novembre de 8h30 à 16h00. Les insignes d'identification vous seront remis à ce moment.

Les présentations orales auront lieu dans la salle S1-139, alors que les affiches seront exposées à la Mezzanine (à l'étage au-dessus).

The registration desk is located near the room S1-139 in the Jean-Coutu Building (2940 Chemin de la polytechnique), and open on November 5, from 8:30 am to 4:00.

Oral presentations will take place in Room S1-139 and posters will be displayed in the Mezzanine (level above).

Pauses santé, lunchs et cocktail / Coffee breaks, lunches and cocktail

Les pauses santé, le lunch et le cocktail seront servis à la Mezzanine.

Coffee breaks, lunch and cocktail will be served in the Mezzanine.

ROBERT CEDERGREN COLLOQUIUM IN BIOINFORMATICS 2009

November 5		
	Time	Oral Presentations (room S1-139, Jean-Coutu Bldg.)
Session 1 / Chair : François Major	9:15	Colloquium opening : **Gertraud Burger**, Biochemistry, UdeM, Colloquium organizer
	9:30	**Keynote 1 : John Quackenbush :: Harvard School of Public Health** ***Network and state space models : science and science fiction approaches to cell fate predictions***
	10:30	Coffee break (Mezzanine)
	11:00	M.Sc. 1 : Sébastien Boisvert :: Université Laval *OpenAssembler: assembly of reads from a mix of high-throughput sequencing technologies*
	11:30	M.Sc. 2 : Sandie Reatha :: University of Ottawa *Predicting emerging influenza strains: the weight of the past*
	12:00	M.Sc. 3 : Elenie Godzaridis :: Université Laval *Reinventing peptidoglycan synthesis inhibitors through rational design*
12:30		Lunch + poster session (Mezzanine)
Session 2 / Chair : B. Franz Lang	13:30	Ph.D. 1 : Claudia L. Kleinman :: Université de Montréal *Protein structure representations for evolutionary analysis*
	14:00	Ph.D. 2 : Xuemei Luo :: University of Carleton *Computational approaches towards the design of pools for in vitro selection of complex aptamers*
	14:30	Ph.D. 3: Ethan Kim :: McGill University *Predicting direct protein interactions from affinity purification mass spectrometry data*
	15:00	Coffee break + poster session (Mezzanine)
	15:30	Ph.D. 4 : Béatrice Roure :: Université de Montréal *Qualitative heterogeneity in the evolutionary process and its impact on the phylogenetic inference*
	16:00	Ph.D. 5 : Hamed Shateri Najafabadi :: McGill University *A universal approach for identification of co-occurring features within biological networks*
	16:30	**Keynote 2 : Gary Stormo :: School of Med. Washington Univ. in St.Louis** ***Modeling DNA and RNA regulatory sites and their interactions with proteins***
	17:00	Free time

ROBERT CEDERGREN COLLOQUIUM IN BIOINFORMATICS 2009

November 6		
Time		Oral Presentations (room S1-139, Jean-Coutu Bldg.)
Session 3/ Chair: Mathieu Blanchette	9:00	**Keynote 3 : Eran Segal :: The Weizmann Institute of Science** ***Reading the genome: from DNA sequence to expression***
	10:00	Ph.D. 6 : Dennis Wong :: Dalhousie University *A phylogenomic and metagebolic analysis of enhanced biological phosphorous removal community metagenomes*
	10:30	Coffee break (Mezzanine)
	11:00	Ph.D. 7 : Glenn Hickey :: McGill University *A practical algorithm for estimation of the maximum likelihood ancestral reconstruction expected error*
	11:30	Ph.D. 8 : Norman J MacDonald :: Dalhousie University *Mitigating the effect of shared ancestry in genotype-phenotype association problems with conditional mutual information*
	12:00	Ph.D. 9 : Adriana Munoz :: University of Ottawa *Rearrangement phylogeny of genomes in contig form*
12:30		Lunch + poster session (Mezzanine)
Session 4/Chair:Raphael Gottardo	13:30	Post-doc. 1 : Kelil Abdellali :: Université de Montréal *ALIGNER: A novel algorithm for detecting and aligning related protein sequences*
	14:00	Post-doc. 2 : Robert Flight :: Dalhousie University *Normalization methods for time-course DNA microarray data*
	14:30	Post-doc 3 : Emmanuel Levy :: Université de Montréal *Are all protein-protein interactions functional? Lessons from evolution*
	15:00	Coffee break + poster session (Mezzanine)
	15:30	**Keynote 4 : Laxmi Parida :: Watson Research Center IBM** ***Graph theoretic approaches to RECOMBINOMICS***
16 :30		Awards : **Christian Baron**, Head of Biochemistry Dept. & **Patrice Marcotte**, Head of Computer Science Dept. (S1-139)
16 :50		Colloquium closing : **Gertraud Burger**, Organizer (UdeM) (S1-139)
17:00		Cocktail (Mezzanine)

KEYNOTES

John Quackenbush
Dana-Farber Cancer Institute and the Harvard School of Public Health , Boston, Massachusetts
<johnq@jimmy.harvard.edu>

Network and state space models: science and science fiction approaches to cell fate predictions

Two trends are driving innovation and discovery in biological sciences: technologies that allow holistic surveys of genes, proteins, and metabolites and a realization that biological processes are driven by complex networks of interacting biological molecules. However, there is a gap between the gene lists emerging from genome sequencing projects and the network diagrams that are essential if we are to understand the link between genotype and phenotype. 'Omic technologies such as DNA microarrays were once heralded as providing a window into those networks, but so far their success has been limited, in large part because the high-dimensional they produce cannot be fully constrained by the limited number of measurements and in part because the data themselves represent only a small part of the complete story. To circumvent these limitations, we have developed methods that combine 'omic data with other sources of information in an effort to leverage, more completely, the compendium of information that we have been able to amass. Here we will present a number of approaches we have developed, including an integrated database that collects clinical, research, and public domain data and synthesizes it to drive discovery and an application of seeded Bayesian Network analysis applied to gene expression data that deduces predictive models of network response. Looking forward, we will examine more abstract state-space models that may have potential to lead us to a more general predictive, theoretical biology.

Gary Stormo
Division of Biology and biomedical sciences, School of Medicine Washington, St-Louis, Missouri
<stormo@wustl.edu>

Modeling DNA and RNA regulatory sites and their interactions with proteins

New sequencing methods, and other technological advances, provide abundant data for studying the interactions of proteins with DNA and RNA and how those interactions are used in the regulation of gene expression. This talk will cover some of our work on modeling regulatory motifs and transcriptional and post-transcriptional regulatory networks.

Eral Segal
Department of Computer Science and Applied Mathematics, The Weizmann Institute of Science, Rehovot, Israel <eran.segal@weizmann.ac.il>

Reading the genome: from DNA sequence to expression

Complex transcriptional behaviors are encoded in the DNA sequences of gene regulatory regions. Advances in our understanding of these behaviors have been gained recently by quantitative models that describe how molecules such as transcription factors and nucleosomes interact with the genomic sequence. An emerging view is that every regulatory sequence is

associated with a unique binding affinity landscape for each molecule and consequently, with a unique set of molecule binding configurations and transcriptional outputs. I will present a quantitative framework based on the hypothesis of competitive binding equilibrium that unifies these ideas, and show that it explains many experimental observations regarding binding patterns of factors and nucleosomes, and dynamics of transcriptional activation. The framework can also be used to model more complex phenomena such as transcriptional noise and the evolution of transcriptional regulation.

Laxmi Parida
Watson Research Center, IBM, Yorktown Heights, New York
parida@us.ibm.com

Graph theoretic approaches to RECOMBINOMICS

The work is motivated by the need for understanding, and processing, the manifestations of recombination events in chromosome sequences. It turns out that one of the tools of choice is graph-theory, which provides a convenient handle on the problems arising in recombinational population genomics. In this talk, we focus on two related problems. First, we explore the general problem of reconstructability of pedigree history. How plausible is it to unravel the history of a complete unit (chromosome) of inheritance? We use a random graphs framework to study pedigree history in an ideal (Wright Fisher) population. This framework correlates the underlying mathematical objects in pedigree graph, mtDNA or NRY Chromosome tree, ARG (Ancestral Recombinations Graph) etc. used in population genomics literature, into a single unified random graph framework. Apart from providing natural and topology-based definitions for many population genomic entities (such as GMRCA and ARG), the framework also suggests sampling algorithms (with proven randomness) to construct random populations for simulations studies. Finally, using a suitable measure, the framework provides a concrete answer to the general reconstructability question.

The second problem deals with estimating the recombinational history of a sample of individuals. Apparently there is more genetic diversity in human genomes, in terms of SNPs, inversions, copy number variations and so on, than was believed earlier. As large databases become available, is it possible to detect the recombination events in the data. And, what stories, if any, these recombinational landscapes tell us? I will discuss our experience with designing discovery algorithms-- a graph based system called IRiS that we have developed for the estimation purposes.

I will conclude with a discussion on our ongoing work in the Genographic Project on the study of human population diversity based on evidence of past recombinations (termed recotypes) as genetic markers. The inferred recombinations indicate strong agreement with past in vitro and in silico recombination rate estimates. The correlation between traditional allele frequency based distances and recombinational distances bring further credence to the study of population structure using recotypes. Also, we make the surprising observation that recotypes are more representative of the underlying population structure than the haplotypes they are derived from.

PRÉSENTATIONS ORALES / ORAL PRESENTATIONS

M.Sc. 1 : Sébastien Boisvert :: Université Laval, Québec, Québec

OPENASSEMBLER : ASSEMBLY OF READS FROM A MIX OF HIGH-THROUGHPUT SEQUENCING TECHNOLOGIES

An accurate and complete genome sequence of a desired species or phylogenetically close relative is now a basic pre-requisite for advanced genomics research. A crucial step in obtaining high-quality genome sequence is the ability to correctly assemble short individual sequence reads into longer contiguous sequences accurately representing genomic regions that are much longer than any single contributing read. Current sequencing technologies continue to offer increases in throughput and corresponding reductions in cost and time. Unfortunately, the benefit of obtaining very large numbers of reads is complicated by a non-trivial presence of sequence errors, with different types of errors and biases being observed with the different sequencing systems. Although software systems exist for assembling reads for each individual system, no comprehensive procedure was proposed for high-quality genome assembly based on mixes of reads from different technologies. We describe an open source software program called OpenAssembler which has been specifically developed to assemble reads obtained from a combination of sequencing systems, and compare its performance to other assembly packages on simulated and real datasets. To illustrate the value of OpenAssembler, we used a combination of Roche/454 and Illumina reads to assemble the 3.6 Mb Acinetobacter baylyi ADP1 genome (NCBI/Genbank accession CR543861) into 119 contigs containing 26 mismatches and 7 indels. The Newbler assembler, using only the Roche/454 reads (reads for which it has been design for), assembled the genome into 118 contigs with 64 mismatches and 356 indels.

M.Sc. 2 : Sandie Reatha :: University of Ottawa, Ottawa, Ontario

PREDICTING EMERGING INFLUENZA STRAINS: THE WEIGHT OF THE PAST

Candidate strains used in the influenza vaccine are chosen in an empirical fashion, and successful selection of strains for the vaccine is critical to human health. In this regard, we previously developed a posterior predictive model that generates emerging sequences conditional on currently circulating strains. However, it is unclear how far back in time sequences need to be sampled in order to maximize predictive power. Here we assess the predictive power of our model for several sampling durations, and show that the best predictive power is obtained for the longest time interval that does not encompass an antigenic shift when predicting the emergence of a new H3N2 variant in the 2007-2008 influenza season.

M.Sc. 3 : Elenie Godzaridis :: Université Laval, Québec, Québec

REINVENTING PEPTIDOGLYCAN SYNTHESIS INHIBITORS THROUGH RATIONAL DESIGN

In Gram-negative bacteria, the cell wall peptidoglycan synthesis cytosolic pathway includes six essential enzymes, MurA, MurB, and the four amide ligases MurC to MurF. These enzymes convert UDP-N-acetyl-glucosamine into UDP-N-acetyl-muramoyl-pentapeptide, the last cytosolic intermediate in peptidoglycan synthesis, and as such, failure of any enzyme in the pathway gives a lethal phenotype. Their primary role as well as their conserved structure and, especially for the ligases (MurC to MurF), their peptidic substrates make them promising drug targets.

In a series of preliminary experiments using purified MurA to MurF enzymes from P. aeruginosa and phage display screening, a method for fast biochemical enzyme binding ability assay, we identified a collection of “lead compound” peptide inhibitors from two libraries: 12-mer (12-residue linear compounds) and C-7-C (compounds circularized by a disulfide bond,

seven residues in length). The lead compounds presented here are referred to as PEP1328, murDp1 and murFp1 for murB, D and F respectively, and have low IC50 and Ki values, in the micromolar range, of 66uM, 4uM and 300uM, respectively.
Bioinformatic analyses were used in this study to aid visualization and characterization of the interplay between enzyme and peptide leading to the results yielded by phage display and combinatorial chemistry screening, and to expose atomic parameters involved in inhibitor binding and activity. We present computer-generated three-dimensional models built upon E. coli crystal structure of MurB, MurD and MurF resolved at 1.80A, 1.52A and 2.30A, closing in on the inhibitory conformations.
Further analyses by comparison with known peptides and biologically active homologs from databases identified structurally closely related compounds such as the peptide tigerinin-1, a circular amphibian antimicrobial. These in silico observations hinted at possible refinements in combinatorial synthesis -- chemical and structural alterations to the peptides. Notably, circularization preserves the active conformation thus decreasing the prevalence of the inactive conformers, while homology to tigerinin-1 could help murDp1 penetrate the external bacterial membrane by mimicking the tigerinins' hydrophobicity. As membrane penetration is a problematic area for any peptidic antibacterial compound, the analyses point to ways to improve cellular uptake without compromising the inhibitory conformation shown in the Mur enzymes. Such synthetic alternatives could enhance lead compounds' effectiveness and suitability as antibacterial peptides targeting Gram-negative bacteria such as Pseudomonas aeruginosa, whose high resistance to antibiotics is infamous.

Ph.D. 1 : Claudia Laura Kleinman :: Université de Montréal, Montréal, Québec

PROTEIN STRUCTURE REPRESENTATIONS FOR EVOLUTIONARY ANALYSIS

The influence of three dimensional protein structure on sequence evolution has not been extensively characterized. The main limitation has been the computationally justified assumption of independence between sites made by probabilistic models in phylogenetics. Recently, models that include an explicit treatment of protein structure and site interdependencies have been developed: a statistical potential (an energy-like scoring system for sequence-structure compatibility) is used to evaluate the probability of fixation of a given mutation. Yet, due to the novelty of these models and the little overlap between the fields of structural and evolutionary biology, only very simple representations of the protein structure have been used so far. Here, we present new forms of statistical potentials, using a probabilistic framework recently developed for the explorations in sequence space involved in evolutionary studies. Terms related to pairwise distance interactions, torsion angles, solvent accessibility and flexibility of the residues are included in the potentials, so as to study the effects of the main factors known to influence protein structure. The new potentials, with a more detailed representation of the protein structure, yield a better fit. When included into a structurally constrained codon model of sequence evolution, they lead, as expected, to an improvement in the description of the evolutionary process. Contrasted to sophisticated site-independent models, however, they still leave a large fraction of the selective constraints unexplained. Altogether, the framework presented here for designing and evaluating sequence-structure compatibility criteria should allow for large-scale studies, to better understand the effect of protein structure on sequence evolution.

Ph.D. 2 : Xuemei Luo :: University of Carleton, Ottawa, Ontario

COMPUTATIONAL APPROACHES TOWARDS THE DESIGN OF POOLS FOR IN VITRO SELECTION OF COMPLEX APTAMERS

Random pools that are used in in vitro aptamer selection are not structurally diverse. Simple topological structures such as stem-loops or low complexity structures dominate in these pools. This lack of structural diversity in random pools explains why complex structure motifs with high-order junctions are rare in selected aptamers. Recent experimental findings suggest that increased structural diversity of starting RNA/DNA pools can enhance the possibility of finding novel aptamers with improved activities. In this study, two computational methods were developed to improve the design of RNA/DNA pools. The Random Filtering method selectively increased the number of high complex structures such as five-way junctions in RNA/DNA pools, whereas the Genetic Filtering method allowed for the design of RNA/DNA pools with uniform structure distribution (i.e. 20 percent 1-way, 2-way, 3-way, 4-way and 5-way junctions each). Biological experimental results confirmed that these methods greatly improved probabilities to access high complex structures in in vitro selection experiments.

Ph.D. 3 : Ethan Kim :: McGill University, Montréal, Québec

PREDICTING DIRECT PROTEIN INTERACTIONS FROM AFFINITY PURIFICATION MASS SPECTROMETRY DATA

Affinity purification followed by mass spectrometry identification (AP-MS) is an increasingly popular approach to observe protein-protein interactions (PPI) in vivo. One drawback of AP-MS, however, is that it is prone to detecting indirect interactions. In this paper, we give a simple mathematical model for separating direct interactions from indirect ones. Given idealized quantitative AP-MS data, we consider the problem of identifying the most likely set of direct interactions that may have produced the observed data. We address this challenging graph theoretical problem by first characterizing signatures for weakly connected vertices as well as dense regions of the network. The rest of the direct PPI graph is then inferred using a genetic algorithm. The accuracy of the algorithm is assessed on both simulated data and idealized biological networks. Then the algorithm is used to predict direct interactions from a large set of AP-MS PPI data for Saccharomyces cerevisiae (yeast).

Ph.D. 4 : Béatrice Roure :: Université de Montréal, Montréal, Québec

QUALITATIVE HETEROGENEITY IN THE EVOLUTIONARY PROCESS AND ITS IMPACT ON THE PHYLOGENETIC INFERENCE

Model violations constitute probably the major limitation in inferring accurate phylogenies. Characterizing which properties of the data are not being correctly handled is therefore of prime importance. One of the characteristics of protein evolution is the variation of evolutionary rate across time, a phenomenon called heterotachy. Its effect on phylogenetic inference has recently obtained considerable attention, which led to the development of new models of sequences evolution. However, this focus on the quantitative heterogeneity of the evolutionary process overlooks the qualitative one. Here, we looked for the importance of variation of the amino-acid substitutional process across time and its possible impact on the phylogenetic inference. We used the CAT model [MBE 2004, 6:1095-109] to define the substitution profiles defined by their equilibrium frequencies over the twenty amino acids, a powerful proxy for qualitatively characterizing the evolutionary process. Using two large datasets, we show that qualitative changes in substitution properties across time occurred significantly more than expected by chance. No significant correlation between sites showing a change in substitution profiles and heterotachous sites is found, suggesting that variation in the amino-acid replacement process is not a consequence of rate change. To test whether this non-

modeled variation can lead to an erroneous phylogenetic tree, we analyzed a concatenation of mitochondria encoded proteins for which Cnidaria and Porifera were erroneously grouped. The progressive removal of the sites with the most heterogeneous profiles allows to recover the monophyly of Eumetazoa (Cnidaria+Bilateria) showing that this heterogeneity can influence the phylogenetic inference by introducing an artifact. The time-heterogeneity of the amino-acid replacement process is therefore an important evolutionary process that should be incorporated in future models of sequence evolution.

Ph.D. 5 : Hamed Shateri Najafabadi :: McGill University, Montréal, Québec

A UNIVERSAL APPROACH FOR IDENTIFICATION OF CO-OCCURING FEATURES WITHIN BIOLOGICAL NETWORKS

Biological networks connect related genes together based on their physical interactions, functional relationships, co-expression, genetic interactions, etc. Such networks are often enriched with pairs of co-occurring gene features, i.e., pairs of features that occur mostly among interacting genes. These features may include biological functions, protein domains, short protein motifs, expression patterns, regulatory elements, phylogenetic profiles, etc. We have devised a method, called FICoPE, which utilizes an information theory-based approach for Finding Informative Co-occurring Pairs of Elements. FICoPE can be used for identification of functional and physical relationships among a wide variety of predetermined features, as well as for de novo identification of co-occurring motifs in either protein or nucleic acid sequences. For example, applying FICoPE to a co-expression network of yeast genes, we were able to identify known and novel transcription factor binding sites with very high specificity. Testing FICoPE on several randomly shuffled networks, we estimate that it has a near-zero false discovery rate.

We also analyzed two different yeast protein interaction networks to identify pairs of functionally/physically linked protein domains. Using the same protein interaction networks, FICoPE also identified short protein motifs and established their relationships with each other and with known protein domains, resulting in a network of functionally/physically linked sequence elements. Some of these motifs are known consensus motifs involved in protein-protein interactions and, as expected, many of them show significantly high network-level conservation between yeast and other organisms, strongly suggesting that they are functional sequence elements. In addition to its extremely low false discovery rate, FICoPE is also able to extract efficiently the functional sequence elements from noisy networks. FICoPE could also establish the relationships among Gene Ontology terms within these interaction networks. It also showed that the pattern of presence or absence of proteins in particular organisms can be used as an indicative of their interactions. We could also use FICoPE to confirm our previous finding that genes with similar codon usages tend to have more physical/functional interactions than genes with different codon usages.

Pairs of elements that are found by FICoPE can also be used for building classifiers that are able to predict connections among genes, such as protein-protein interactions. We have devised an algorithm, called PIPE (Predicting Interactions based on Pairs of Elements), which heuristically searches for the best set of element pairs that can predict the connections within a biological network. We have successfully used PIPE to predict protein-protein interactions and protein complexes in yeast with the highest reported specificity and sensitivity. Based on these results, we present FICoPE/PIPE as a universal framework for integrating almost all types of biological data for analysis and prediction of biological networks.

Ph.D. 6 : Dennis Wong :: Dalhousie University, Halifax, Nova Scotia

A PHYLOGENOMIC AND METAGEBOLIC ANALYSIS OF ENHANCED BIOLOGICAL PHOSPHOROUS REMOVAL COMMUNITY METAGENOMES

The wealth of sequenced genomic data and recently, sequencing of genomes from environmental samples (i.e. Metagenomes) has opened the door to the analysis of microbes that were never before possible. However, this new found wealth of data requires modeling approaches and computational methods that can represent the biology and underlying evolutionary processes for the organisms of interest, but at the same time must be computationally efficient. Here I analyse two geographically isolated enhanced biological phosphorous removal (EBPR) communities by comparing their genetic content to fully sequenced reference genomes. I identify unique evolutionary relationships for genes in the EBPR communities using this set of reference genomes. The metabolism for the EBPR communities is then modeled to gain insights into the biology of organisms sampled from an environment that varies dramatically through time.

To identify unique evolutionary relationships, I first obtain clusters of orthologous genes. Many clustering approaches exist, and they span the spectrum of computational efficiency and relevancy to biological data sets. To emphasize biological relevance while still being computationally efficient, I use a divide and conquer approach to construct clusters of orthologous genes from 840 genomes and the two EBPR community metagenomes. Closely related genes are first identified using BLAST and clusters of orthologs are constructed at the taxonomic class level using the triangle equality of best hits. Clusters are joined together to obtain complete taxonomic coverage. From the clusters, I then obtain alignments and construct phylogenies to identify cases of unique evolutionary history.

Metabolism is the set of chemical processes that support growth and reproduction for an organism. To model metabolism, I use the KEGG database and construct networks, where nodes are enzymes, and undirected edges connecting nodes are chemicals shared between enzymes. The shape of metabolic networks provides insights into processes important to the biology of life, especially ecology and evolution when observed across a diverse set of taxa. I examine network properties such as modularity, average node clustering coefficient, and average path length, which indicates the ability to adapt to novel environments, redundancy of metabolism and efficiency of metabolism, respectively. To obtain such statistics, I use the open source program GraphCrunch, Newman's "Fast Modularity" algorithm, and the network visualization tool Cytoscape. I then use randomization procedures to determine if EBPR community members differ from relatives across the calculated network properties.

Results of phylogenetic analyses identify unique evolutionary relationships for a variety of genes important to taxa important in EBPR community function. Examination of the phylogenies suggests the possibility of lateral gene sharing in EBPR communities. Measures of network properties such as modularity reveal that Accumulibacter phosphatis, the primary phosphorous removing bacteria in the EBPR community, tend to have more modular metabolic networks compared to close relatives. This suggests that for communities in an environment that is highly variable, modularity is required to ensure survival and reproduction.

Ph.D. 7 : Glenn Hickey :: McGill University, Montréal, Québec

A PRACTICAL ALGORITH FOR ESTIMATION OF THE MAXIMUM LIKELIHOOD ANCESTRAL RECONSTRUCTION EXPECTED ERROR

The ancestral sequence reconstruction problem asks to predict the DNA or protein sequence of an ancestral species, given the sequences of extant species. Such reconstructions are fundamental to comparative genomics, as they provide information about extant genomes and

the process of evolution that gave rise to them. Arguably the best method for ancestral reconstruction is maximum likelihood estimation. Many effective algorithms for accurately computing the most likely ancestral sequence have been proposed. We consider the less-studied problem of computing the expected reconstruction error of a maximum likelihood reconstruction, given the phylogenetic tree and model of evolution, but not the extant sequences. This situation can arise, for example, when deciding which genomes to sequence for a reconstruction project given a gene-tree phylogeny (The Taxon Selection Problem). In most applications, the reconstruction error is necessarily very small, making Monte Carlo simulations very inefficient for accurate estimation. We present the first practical algorithm for this problem and demonstrate how it can be used to quickly and accurately estimate the reconstruction accuracy. We then use our method as a kernel in a heuristic algorithm for the taxon selection problem.
The implementation is available at http://www.mcb.mcgill.ca/~blanchem/mlerror.

Ph.D. 8 : Norman J. MacDonald :: Dalhousie University, Halifax, Nova Scotia

MITIGATING THE EFFECT OF SHARED ANCESTRY IN GENOTYPE-PHENOTYPE ASSOCIATION PROBLEMS WITH CONDITIONAL MUTUAL INFORMATION

Identifying genotype-phenotype relationships is important for understanding what causes the observed traits of organisms. Simple predictive accuracy between presence and absence of genotypes versus phenotypes can help to identify relationships, however, we can find many spurious relationships if we do not take into account the dependency that occurs among samples due to shared ancestry. Previously, phylogenetic methods have been applied to account for coalescence while looking for associations within HIV. Other studies have used classifiers to find predictive associations between genotype and phenotype, without accounting for common descent. In this study, we propose a novel application of Conditional Mutual Information (CMI) for subtracting the influence of common descent of genes and phenotypes when examining a predicted gene to phenotype relationship.
Classification based on Predictive Association Rules finds rules of the form A->C where the presence of the set A (the antecedent) is predictive of the presence of the set C (the consequent). Classifiers built with associative rules have a significant interpretability advantage over other classification algorithms such as support vector machines, as the results of the training are human readable.
Conditional mutual information (CMI) is defined as the information shared between two random variables in the context of a third. In this case, the two random variables we are interested in are the presence and absence profiles of genes and phenotypes, while the confounding factor is the taxonomic classification of organisms. Organisms in the same group potentially share genes for reasons of common descent rather than functional similarity.
To quantify the evidence we have for a genotype-phenotype association in the context of taxonomy, we use a novel score, the Conditionally Weighted Mutual Information (CWMI). CWMI is defined as the mutual information shared between genotype and phenotype, scaled by the ratio of the CMI to the maximum possible CMI given the taxonomic label distribution. This measure down-weights the accuracy score of associations for which there is evidence of common ancestral descent having a confounding effect, and thus we have little evidence that they are in fact causal of the phenotype. In this study, we have measured if conditionally weighting various measures such as mutual information leads to better generalizability of the results, by testing 365 phylogenetic profiles on 5 binary phenotypes (thermophily, halophily, motility, spore formation, and photosynthesis) using 2-fold stratified cross-validation. In this

presentation, we report the performance on each of the 5 datasets for a variety of accuracies, both accounting for shared descent and not accounting for shared descent.

Ph.D. 9 : Adriana Munoz :: University of Ottawa, Ottawa, Ontario

REARRANGEMENT PHYLOGENY OF GENOMES IN CONTIG FORM

There has been a trend in increasing phylogenetic coverage for genome sequencing while decreasing the sequencing coverage for each genome. With lower coverage, there is an increasing number of genomes being published in contig form. Rearrangement algorithms, including gene order-based phylogenetic tools, require whole genome data on gene order, segment order, or some other marker order. Items whose chromosomal location is unknown cannot be part of the input. The question we address here is, for gene order-based phylogenetic analysis, how can we use rearrangement algorithms to handle genomes available in contig form only? Our suggestion is to use the contigs directly in the rearrangement algorithms as if they were chromosomes, while making a number of corrections, e.g., we correct for the number of extra fusion/fission operations required to make contigs comparable to full assemblies. We model the relationship between contig number and genomic distance, and estimate the parameters of this model using insect genome data. With this model, we can then reconstruct the phylogeny based on genomic distance and numbers of contigs.

Post-doc. 1 : Kelil Abdellali :: Université de Montréal, Montréal, Québec

ALIGNER: A NOVEL ALGORITHM FOR DETECTING AND ALIGNING RELATED PROTEIN SEQUENCES

The alignment of protein sequences is the process of finding the best matching between these sequences by inserting "gaps" in the appropriate positions in each sequence, so that the positions from the sequences with identical or similar residues are aligned. The alignment aims to identify regions of similarity that might reveal significant patterns of functional, structural, or evolutionary significance in a given set of protein sequences. The literature reports two types of alignment approaches, global and local. In one hand, global alignment approaches span the entire length of all protein sequences by aligning every residue in every sequence. On the other hand, local alignment approaches look for most conserved patterns by identifying regions of similarity within long protein sequences that are often widely divergent overall. It has been shown by McClure et al. 1994 and Thompson et al. 1999 that the most effective alignment approaches depend essentially on the structural nature of protein sequences to be aligned. Oftentimes global alignment produces the most accurate and reliable alignments, but in the presence of large N/C-Terminal extensions and internal insertions local alignment is the most successful. This is even more so when it comes to multi-modular protein sequences. The most important problem with these two approaches is that, without prior knowledge about the biochemical and structural properties of each of these proteins, we cannot choose with high certainty the appropriate approach to perform the alignment that can reveal the patterns most relevant functionally or structurally in these sequences. This is the main reason why ALIGNER is proposed.

On the other hand, existing alignment approaches are devised to produce for a given input dataset the alignment of all the protein sequences, and ignore if the input dataset includes divergent protein sequences, those who do not share enough of conserved regions to produce significant alignments. This can complicate the identification of regions of similarity. To deal with this problem, biologists oftentimes handle manually input datasets by discarding protein sequences that may disturb the alignment, which is not always possible especially when input datasets include several divergent groups of protein sequences. This is the second reason why ALIGNER is proposed.

We present sALIGNER here, a new and effective alignment algorithm that is able to align effectively protein sequences that need either global or local alignment. As in global alignment, sALIGNER spans the entire length of all protein sequences by aligning every residue in every sequence. At the same time, ALIGNER gives a particular attention to the significant patterns shared between protein sequences. In addition, ALIGNER detects in input protein datasets, groups of protein sequences that share enough significant patterns to produce alignments that can reveal important structural and functional properties within each group, and this without resorting to user manipulations of the input datasets.

Post-doc. 2 : Robert Flight :: Dalhousie University, Halifax, Nova Scotia

NORMALIZATION METHODS FOR TIME-COURSE DNA MICROARRAY DATA

Normalization is an important step in any DNA microarray analysis, with the goal of removing bias that may interfere with the identification of truly differentially expressed genes. Many different normalization methods have been proposed in the literature, however, the majority are only applicable to experiments wherein two samples are compared to one another (comparator experiments). Time-course DNA microarray experiments (and other serial types of experiments), wherein the arrays form a pseudo-continuum of expression, have become more common in recent years, with no concomitant discussion or development of appropriate normalization methods. We have developed a conceptually simple technique for normalizing DNA time-course data that we refer to as sequential normalization. In this presentation the rationale of sequential normalization will be explained and compared to other frequently used normalization methods using three different time-course DNA microarray data sets. The appropriateness of the commonly employed LOWESS method with respect to time-course DNA microarray experiments will also be examined.

Post-doc. 3 : Emmanuel Levy :: Université de Montréal, Montréal, Québec

ARE ALL PROTEIN-PROTEIN INTERACTIONS FUNCTIONAL? LESSONS FROM EVOLUTION

A paradox stems from the discrepancy between the large numbers of protein-protein interactions (PPIs) characterized by large-scale experiments, and the comparatively smaller number of PPIs that we can make biological sense of. This paradox fuels debates around a fundamental question: what do all these interactions mean? We argue that a large number of physical PPIs may simply be nonfunctional, or noisy. That is, they do exist in cells, they can be detected by typical PPI assays, but they have not evolved to achieve a particular function. Instead, just like spurious transcription factor binding sites, they appear and disappear rapidly during the random course of evolution. Importantly, their existence could explain why PPIs determined from large-scale studies often lack functional relationships between interacting proteins and why the PPI space appears to be immensely large (~20,000 in yeast, and ~130,000 in human). This motivated us to examine the possible existence of noisy interactions as well as some of their implications. First, I will present a formalism that we developed to assess whether noisy interactions are expected to exist. This formalism allowed us to anticipate that noisy interactions should indeed exist. Second, the existence of a large number of noisy interactions implies that they can appear rapidly during the course of evolution, i.e., that a single or a few point mutations can significantly affect the stickiness between proteins. This hypothesis led us to show that "dangerous" mutations, which are likely to induce stickiness with other proteins, are selected against. Finally, because they are assumed to be non-functional, noisy interactions should not be conserved across organisms. This idea prompted us to assess the conservation of a particular type of PPI: kinase-substrate interactions, and predict

that the fraction of functional phosphorylated sites appears in fact smaller than the fraction of non-functional ones.
All three parts of our analyses point to the idea that noisy interactions do exist in cells. This is an important issue to consider because their existence could contribute to explain why PPIs determined from large-scale studies often lack functional relationships between interacting proteins, why PPIs are poorly conserved across organisms, and why the PPI space appears to be immensely large. Our results also stress that comparative proteomics approaches between closely related organisms should be used for the study of protein interactions.

AFFICHES / POSTERS

M.Sc. 1 : Robert Eveleigh :: Dalhousie University, Halifax, Nova Scotia

BEING AQUIFEX AEOLICUS: UNTANGLING A HYPERTHERMOPHILE'S CHECKERED PAST

The hyperthermophilic bacterium Aquifex aeolicus and its sister lineages often appear as an early-branching 'lonely lineage' in trees of 16S rDNA and genome phylogenies. However, the true position of the Aquificae group remains controversial, with extensive lateral gene transfer and thermophilic adaptations further clouding an already unclear organismal lineage. This group has a strong affinity with other groups, notably the (epsilon) Proteobacteria and certain Archaea, but previous work based on gene trees and concatenates has nonetheless been interpreted to support an early-branching position for this group. Here we use a combination of phylogenetic profile- and tree-based approaches to support an alternative explanation, namely that Aquifex aeolicus and its close relatives are sisters to the epsilon-Proteobacteria. These observations are in agreement with potentially informative properties that are shared by these two groups, and demonstrate the remarkable evolutionary plasticity of these lineages.

M.Sc. 2 : Elenie Godzaridis :: Université Laval, Québec, Québec

REINVENTING PEPTIDOGLYCAN SYNTHESIS INHIBITORS THROUGH RATIONAL DESIGN

(See oral presentation abstract)

M.Sc. 3 : Jean-Christophe Grenier :: Université de Montréal, Montréal, Québec

PHYLOGENOMICS OF ARCHAEA

Despite the recent acceleration in sequencing of new genomes of prokaryotes, our knowledge about their phylogeny and divergence times has not improved accordingly. At this time, phylogenies based on 16S ribosomal RNA are still largely the reference system. The analysis of large datasets has been limited, mainly due to the unpredictable impact of horizontal gene transfer (HGT) on these phylogenies. Altogether, these problems slow down considerably our understanding of prokaryotic phylogeny. So far, only a relatively small number of phylogenomic studies of Archaea were published, but the phylogenetic trees obtained usually lack statistical support. Starting our study with the less numerous Archaea, we will use methods that were successfully developed for eukaryotes to generate large datasets, correcting for systematic errors that could affect the analysis. We want to test the hypothesis that the non-phylogenetic signal, mainly a result of systematic errors generated by the incapacity of most current models of sequence evolution to correctly account for properties of the data (e.g. compositional bias or covarion structure), largely explains the low resolution currently observed within the Archaea. We will test if the resolution increases when using both larger datasets and better methods that correct these systematic errors. Here, we present a protocol to efficiently build a large high-quality dataset of orthologous genes. First, sets of putative orthologous genes are constructed with OrthoMCL, which is based on BLAST similarity score. We subsequently concatenate these putative orthologous sets and calculate a matrix representing the pairwise distance between species. This matrix is very robust to orthology errors (paralogy and xenology that are not detected) and constitute a good reference to detect non-orthologous or fast evolving sequences for individual genes. We applied this protocol to 79 Archaea. The preliminary phylogenies obtained are promising and, if confirmed, will have an impact on the currently assumed position of the Thermoplasmatales group. Our analysis puts this group closer to the base of Euryarchaeota, while the previously published analyses put

them among the methanogens. This could mean that the origin of methanogenesis within the Euryarchaeota is more recent than currently assumed.

1. Rodriguez-Ezpelata et al. (2007) Detecting and Overcoming Systematic Errors in Genome-Scale Phylogenies, Syst. Biol., 56:3, 389-99.
2. Brochier-Armanet et al. (2008) Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota, Nature Reviews, 6:245-52.
3. Li et al. (2003) OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes, Gen. Res., 13:2178-89.

M.Sc. 4 : François Lefebvre :: Université de Montréal, Montréal, Québec

CO-REGULATION AS A BIOLOGICAL BENCHMARK TO UNCOVER DIFFERENTIAL EXPRESSION

Microrrays remain an important tool for the measurement of gene expression, and a myriad of methods for their analysis has been proposed in the past. However, insufficient and sometimes contradictory evidence has prevented the emergence of a consensus, thus leaving microarray practitioners to somewhat arbitrarily decide which method should be used to analyze their data. Here we present a novel approach to the problem of comparing methods for the identification of differentially expressed genes. Over eight hundred analytic pipelines were applied to 87 independent microarray experiments. The accuracy of each analytic pipeline was assessed by measuring the average level of co-regulation uncovered across all data sets. This analysis thus relies on a varied set of biologically relevant data, does not confound reproducibility for accuracy and can easily be extended to future analytic pipeline. This procedure identified FARMS summarization and the TREAT gene ordering statistic as algorithms significantly more accurate than other alternatives. Based on simulation data, we exhaustively explored the parameter space of regularized statistics and identified appropriate directions to increase the accuracy of current regularized statistics well above what is achieved by fold-change alone. Our results independently confirm the MAQC conclusion regarding the superiority of fold-change when compared to the traditional T-test or regularized statistics, but clearly demonstrates that superior statistics exist.

M.Sc. 5 : Seyyid Ahmed Medjahed :: Université de Mostaganem, Mostaganem, Algérie

L'INDEXATION ET LA RECHERCHE D'INFORMATION PAR ONTOLOGIE DE DOMAINE, DOMAINE DE MEDECINE

On a créé un moteur d'indexation et recherche d'information par ontologie, ontologie de medecine. L'application permet d'indexer les documents de types : PDF-Doc-Rtf-TXT-Html, et permet aussi d'indexer une page web en connaissant son URL. Pour les documents locaux sur le disque, le système va charger le document et charger l'ontologie, il calcule le pourcentage des concepts existants dans le document, si le pourcentage> à 10, alors on peut dire que le document est un document de médecine sinon ce n'est pas un document de médecine. Pour une page html on doit passer par l'étape de nettoyage qui consiste éliminer les balises avant d'appliquer l'algorithme. Si le document est un document de médecine alors on va effectuer une 2ème classification, pour préciser le type de maladie qu'il traite. Pour chaque maladie on a calculé la somme des poids de tous les concepts qui existent dans le document, et on va opter pour la somme maximale. Pour valider notre algorithme, on a rassemblé 1000 textes, 500 de médecine et 500 de non médecine, après avoir fait nos tests on a eu un taux de succès qui avoisine les 92%. Et pour finir, on a intégré dans notre application un système qui permet de compléter et d'enrichir notre ontologie, en intégrant de nouvelles maladies avec et tous ses concepts et de générer la carte conceptuelle pour la représenter graphiquement.

M.Sc. 6 : Eloi Mercier :: Université de Montréal, Montréal, Québec

A BIOCONDUCTOR PIPELINE FOR THE ANALYIS OF CHIP-CHIP AND CHIP-SEQ EXPERIMENTS

Chromatin immunoprecipitation followed by either genome tiling array analysis (ChIP-on-chip) or massively parallel sequencing (ChIP-seq) enables transcriptional regulation to be studied on a genome-wide scale. By systematically identifying protein-DNA interactions of interest, studies using these technologies provide information on cis-regulatory circuitry underlying various cellular processes. However, analysis of the massive and heterogeneous datasets from these studies poses several challenges, including effective data visualization, seamless connection of low-level (close to raw data) and high-level (close to answering biological questions) analysis, integration of data from multiple technological platforms, and flexibility to customize the analysis so that specific biological questions cans be addressed. Although there are several recently developed programs that target some of the individual steps, we want to develop some different integrated R packages that can satisfy all basic needs in ChIP data analyses. We have developed a set of methods in R language to meet these need. We propose to present our bioinformatics/Bioconductor pipeline on the analyse of ChIP-on-chip and ChIP-Seq data. We demonstrate the use of our pipeline by comparative analysis of ChIP-Seq data for the transcription factor FOX-A1 data and we have identified the binding motif for the FOS proto-oncogene protein.

M.Sc. 7 : Raphael Poujol :: Université de Montréal, Montréal, Québec

ESTIMATING PHYLOGENETIC CORRELATIONS BETWEEN MOLECULAR DATA AND LIFE HISTORY TRAITS

Studies on aging suggest that it is due to the accumulation of biochemical damage in DNA, proteins and lipids. Many genes have been proposed to play a role in prevention of cell degeneration, oxidative stress and premature aging. Assuming that these genes are subject to stronger selective pressure in long-lived species, our laboratory use Bayesian modeling to reconstruct the history of longevity and the selective pressure throughout the lineages.

The selective pressure on a gene can be estimated by ω, the ratio of non-synonymous (dN) to synonymous (dS) substitution rates over time (ω = dN / dS). Lower values of ω indicate a stronger selective pressure.

Therefore, when ω is negatively correlated with longevity (i.e. this gene has been under more intense purifying selection in long-living species) the gene is likely involved in the regulation of aging. The estimate of the correlation should be made in a phylogenetic framework, in order to dissociate the dependencies due to evolutionary inertia.

The main idea of my study is to reconstruct the correlated history of longevity and selective pressure (ω) along the lineages of a phylogenetic tree, using a bivariate Brownian process along the phylogeny. The covariance and all the parameters of the model are estimated in a Bayesian MCMC (Markov Chain Monte Carlo) framework using comparative data.

The model has been applied to multiple alignments of candidate genes over 25 mammalians species, allowing the estimation of the posterior probability of a negative correlation between longevity and history of selective pressure. It can be extended to more than two characters so as to address further questions about the interdependence between molecular evolution and life traits (mass, metabolism) or environmental factors (temperature, oxygen).

Ph.D. 1 : Dunarel Badescu :: Université du Québec à Montréal, Montréal, Québec

IDENTIFICATION OF SPECIFIC GENOMIC REGIONS UNDER PRIOR EPIDEMIOLOGIC KNOWLEDGE OF ARCINOGENICITY OR INVASIVITY AND APPLICATION TO HUMAN PAPILLOMA VIRUS AND NEISSERIA MENINGITIDIS

In this poster, we present an algorithm for analyzing the information content of multiple sequence alignments in relation to epidemiologic carcinogenicity or invasivity data to identify regions that would warrant additional experimental analyses. This algorithm is based on a sliding window procedure and a p-value computation to identify genomic regions that are specific to Human Papilloma Viruses (HPV) and Neisseria meningitidis strains causing disease. HPV have a well known potential to cause cervical cancer, while Neisseria Meningitidis is a major causal agent of meningitis and septicaemia worldwide. We present four distance-based discrimination functions for the identification of relevant genomic segments that distinguish between two groups of data. They permit to identify relevant regions and known molecular features. We found that one of the tested functions is specifically well correlated with surface-exposed loops of Neisseria meningitidis outer membrane proteins, regions important in vaccine design.

Ph.D. 2 : Amandine Bemmo :: McGill University, Montréal, Québec

GENOME-WIDE INVESTIGATION OF CHANGES IN PRE-MRNA SPLICING ASSOCIATED WITH METASTASIS OF BREAST CANCER

Breast cancer, the most common cancer among women, is the second leading cause of cancer deaths in women today. Progression of cancer is generally associated with abnormalities in gene splicing and expression. Significant efforts are being made to identify oncogenes based on their expression patterns. Much more poorly understood, but perhaps equally important cancer-related changes, take place at the mRNA processing and, in particular, pre-mRNA splicing levels. Numerous aberrant splice variants appear during the progression of cancer, and the appearance of some of them has been associated with metastasis. We used a new technology, the Affymetrix Exon Array, to carry out concurrent profiling of pre-mRNA splicing and expression at the whole genome level, in the context of changes occurring during the progression of breast cancer. We utilized a well characterized series of three mammary tumor lines exhibiting varying levels of metastatic potential; all possess the ability to form primary mammary tumors when injected into Balb/c mice, but display different abilities to metastasize from the primary site: 168FARN cells are weakly detected in lymph node but fails to cause extravasation; 4T07 cells reach the lung via the blood but are unable to develop metastatic nodules; 4T1, the most metastatic, spontaneously metastasizes to distant sites, namely lung, bone and liver, by the formation of visible nodules in these organs. Statistical analysis identified significant expression changes in 10744 out of 493710 (2%) exon probesets belonging to 2623 out of 16654 (16%) genes, corresponding to putative alternative isoforms that are differentially expressed across tumors of varying metastatic potential. A gene pathway analysis showed that 1224 of these genes have been reported to be involved in diseases and have biological functions predominantly related to cancer, cell interactions, cell proliferation, cell migration and cell death. Our analysis suggests that a large number of genes that exhibit alternative splicing or other isoform changes are associated with metastasis and that these changes may be functionally involved in the progression of cancer. Compared to other approaches based on DNA microarrays that interrogate single whole genes, studying genome wide analysis of alternative splicing in breast cancer at the exon level adds more knowledge about the type of variations occurring in genes; this can easily lead to improved and more specific diagnostics or treatment methods. The detection of ASEs, especially novel ASEs, is of a great importance for

breast-cancer studies. For example, by establishing which genes actively participate in different cancer stages, tumor-specific alternatively spliced mRNAs can be used as breast cancer biomarkers.

Ph.D. 3 : Marc-Frédérick Blanchet :: Université de Montréal, Montréal, Québec

IDENTIFICATION OF 2D MOTIFS THAT INTERACT AND STABILIZE RNA 3D STRUCTURE

The crystal structure of the P4-P6 domain of the T. thermophila group I intron is characterized by the presence of a GNRA tetraloop, which in the 3D structure interacts with a receptor UA_Handle. Both 2D and resulting 3D motifs define nucleotide cyclic motifs (NCM). Both 2D motifs can be identified in the MC-Fold predictions. One way to exploit the 3D interaction network information in 3D folding is to find a building order that includes the 2D as well as the 3D motifs. Here, we developed an MC-Sym input script to model the P4-P6 domain including the tertiary NCM resulting from the GNRA and UA_Handle motifs. This script produces the best P4-P6 model we ever generated, which share 5.8 Å of RMSD with the crystal structure (all-atoms but H). We also present the results of building a database of 2D motifs that interact in 3D. We store the ground and induced versions of the 2D motifs, i.e. those we expect to find in 2D structure predictions and their structures induced by the interactions, and the resulting 3D motifs. This information suffices to identify the possible 3D interaction networks, where each one represents a structural hypothesis inferred from sequence. We will next automate the generation of MC-Sym input scripts that account for this 3D information.

Ph.D. 4 : Alix Boc :: Université du Québec à Montréal, Montréal, Québec

INFERRING AND VALIDATING HORIZONTAL GENE TRANSFER EVENTS USING BIPARTITION DISSIMILARITY

Horizontal gene transfer (HGT) is one of the main mechanisms driving the evolution of microorganisms. Its accurate identification is one of the major challenges posed by reticulate evolution. Here we will present a new polynomial-time algorithm for inferring HGT events and compare three existing and one new tree comparison measures in the context of HGT identification. The proposed algorithm can rely on different optimization criteria, including least-squares (LS), Robinson and Foulds (RF) distance, quartet distance (QD) and bipartition dissimilarity (BD), while searching for an optimal scenario of SPR (Subtree Prune and Regraft) moves needed to transform the given species tree into the given gene tree. As the simulation results suggest, the algorithmic strategy based on BD generally provides better results than the strategies based on the LS function and RF or QD distances. The BD-based algorithm also proved to be more accurate and fast than a well-know polynomial time heuristic RIATA-HGT. Moreover, the HGT recovery results yielded by BD were generally equivalent to those provided by the exponential-time algorithm LatTrans, but a clear gain in running time was obtained using the new algorithm. Finally, a statistical framework for assessing the reliability of obtained HGTs by bootstrap analysis will be also introduced. Some applications of this algorithm will be presented.

Ph.D. 5 : Slim Fourati :: Université de Montréal, Montréal, Québec

PRIMARY ESTROGEN TARGET GENES PREDICT SUCCESS OF TAMOXIFEN THERAPY

Two thirds of breast tumors express estrogen receptor alpha and are candidates for treatment with antiestrogens such as tamoxifen, which has been used in the past 30 years for treatment of all stages of breast cancer. However, tamoxifen treatment is effective only in 50% of cases, and histo-pathological characteristics of the tumors are insufficient to predict tumor evolution upon tamoxifen treatment. Here we examined whether primary estrogen target genes, identified through expression microarray analysis in ER-positive breast carcinoma MCF7 cells (Bourdeau V., 2008), can specifically predict the outcome of antiestrogen therapy, i.e. are indicative of the rate of relapse in tamoxifen-treated but not in untreated tumors. We used a

dataset comprising tumors either treated or non-treated with tamoxifen (Loi S., 2008) to examine the predictive value of primary versus secondary target estrogen genes. A training set of tamoxifen-treated tumors was partitioned into two groups according to levels of expression of primary or secondary estrogen-target genes using a non-supervised approach (k-means partitioning) and a nearest centroïd classifier was used to distribute tumors in the two groups. Kaplan-Meier analysis indicated that a sub-group with statistically significant better prognosis was identified in tamoxifen-treated tumors, but not in tumors not subjected to tamoxifen treatment when using primary target genes. On the other hand, secondary target genes identified a sub-group with better prognosis both in tamoxifen-treated and untreated tumor sets. These results suggest that secondary target genes are of general prognostic value, likely reflecting the proliferative effects initiated by estrogen treatment. On the other hand, the observation that primary estrogen target genes as a group have significant prognostic value only for tamoxifen-treated tumors suggest that they reflect more specifically the mechanisms of action of estrogen and, likely, of antiestrogens used in breast cancer therapy.

Ph.D. 6 : Siyuan Hou :: Dalhousie University, Halifax, Nova Scotia

EXPLORATORY DATA ANALYSIS OF NOISY MICROARRAY DATA WITH MAXIMUM LIKELIHOOD PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is one of the most widely used methods in exploratory data analysis. However, a weakness of PCA is that it assumes homoscedastic errors (uniform measurement error variance), which in real experiments is generally not true. When PCA is used to process data with significantly heteroscedastic errors, its capability will be greatly diminished. In this work, the application of maximum likelihood principal component analysis (MLPCA) in exploratory data analysis was examined. MLPCA was designed to deal with heteroscedastic errors and has been applied in multivariate calibration, parallel factor analysis, and multivariate curve resolution. However, its application in exploratory data analysis is still an area that has not been investigated.

In this work, comparisons between PCA and MLPCA in exploratory data analysis based on simulated data were performed first. It is demonstrated that MLPCA has advantages over PCA in the case of heteroscedastic errors in exploratory data analysis. The advantages of MLPCA over PCA are due to the better estimates of true subspaces and the maximum likelihood projection method. As MLPCA incorporates measurement error information, the error information can also be used to adjust the display of each data point in MLPCA scores plots. A transparency technique was developed which makes the data points with high uncertainties more transparent in scores plots. It is shown that such a technique is helpful to look for useful cluster information in exploratory data analysis.

MLPCA was applied to literature microarray experimental data that studied the gene expression levels of yeast cells exiting out of stationary phase [1]. The data set included the gene expression levels of 19 time points after yeast cells exited out of the quiescent state. The ratios between test channel and reference channel, as well as the ratio uncertainties were estimated by the regression ratio method in the literature [2]. It was found that the data set had significantly heteroscedastic experimental errors and the measurement uncertainties roughly followed a log-normal distribution. Therefore, MLPCA was applied to this data set and the transparency technique was used to help to look for clusters of genes. It was found that genes clustered based on MLPCA scores plots had similar gene expression profiles. Further analysis by searching for gene ontology terms for the extracted genes showed that genes classified by MLPCA have statistically significant association to some gene ontology terms, indicating that the extracted information by MLPCA is biologically meaningful.

1. M. J. Martinez, S. Roy, A. B. Archuletta, P. D. Wentzell, S. S. Anna-Arriola, A. L. Rodriguez, A. D. Aragon, G. A. Quinones, C. Allen, and M. Werner-Washburne. Genomic Analysis of Stationary-Phase and Exit in Saccharomyces cerevisiae: Gene Expression and Identification of Novel Essential Genes. Molecular Biology of the Cell, 2004, 15, 5295-5305.
2. T. K. Karakach, R. M. Flight, and P. D. Wentzell. Bootstrap Method for the Estimation of Measurement Uncertainty in Spotted Dual-Color DNA Microarrays. Analytical and Bioanalytical Chemistry, 2007, 389, 2125-2141.

Ph.D. 7 : Julie Hussin :: Université de Montréal, Montréal, Québec

FINE SCALE PATTERNS OF RECOMBINATION, HAPLOTYPE STRUCTURE AND NATURAL SELECTION IN FRENCH-CANADIAN MULTI-GENERATION PEDIGREES WITH CONGENITAL HEART DISEASE

Population genetic modeling can shed light on processes such as recombination and selective pressures while potentially identifying genes implicated in risk of human disease. Understanding how variation contributes to susceptibility to disease will help characterize genetic variants and establish their role.

We present here an empirical study of French-Canadian multi-generation pedigrees with individuals presenting a left ventricular outflow tract obstruction (LVOTO). Linkage and haplotype analysis of French-Canadian cohorts with LVOTO previously showed high heritability of such lesions and candidate regions have been mapped. We will show in-depth analysis results related to the genetic variation found in 478 individuals from 68 families, genotyped for more than 650,000 SNPs.

We use this genome-wide SNP genotype data to localize crossovers and examine variation in fine-scale recombination patterns among individuals from families with multiple cases. The results are compared with the findings from previous studies of high-resolution crossover mapping in Hutterites families. We further look for signatures of natural selection using novel haplotype allelic classes statistics, that combines information from both segregating sites and haplotype diversity. Reduced haplotype-specific decay of linkage disequilibrium using these same statistics has helped identify potential loci associated with the LVOTO phenotype in the French-Canadian population.

Ph.D. 8 : Louis-Philippe Lemieux-Perreault :: Université de Montréal, Montréal, Québec

CNGEN – A NEW TOOL FOR COPY-NUMBER GENOTYPES PARTITIONING

Copy number polymorphisms (CNPs) and variations (CNVs) may be at least as important as single nucleotide polymorphisms (SNPs) in assessing human genetic variability, since conservative estimates have shown that they might affect more than 10% of the human genome. Integrated genotypes, derived by the genotyping of SNPs, CNPs and CNVs using the Birdsuite software on Affymetrix's Genome-Wide Human SNP array, are now being used for association studies of complex traits. The use of those genotypes in linkage analysis with multi-generational family data is limited by the requirement of chromosome-specific copy number assignment, or partitioning. We have developed new software which, once applied to familial trios or extended pedigrees, produces partitioned copy number genotypes with distinct parental alleles. Those multi-allelic partitioned copy number polymorphisms have the potential to offer a new and powerful tool for linkage analysis. CNGen has been validated using simulations on complex pedigree structures. The simulation steps have shown that CNGen will not result in an excess of false calls in the presence of genotyping error or de novo mutations, supporting its robustness. The new method has been applied successfully to a real dataset of 300 genotyped samples from 42 pedigrees segregating congenital left ventricular outflow tract obstruction. CNGen partitioned 55% of Birdsuite's results and identified 3,500 (0.44%)

Mendelian errors in the process, a rate within expectations for multi-allelic markers. CNGen is a flexible, open source and platform independent Python program.

Ph.D. 9 : Étienne Lord :: Université du Québec à Montréal, Montréal, Québec

ARMADILLO, AN AUTOMATED WORKFLOW SYSTEM FOR PHYLOGENETIC ANALYSIS

Phylogenetic analysis is a necessary task for a researcher studying ancestral genes, orthologs or horizontal gene transfer events (HGT). While a computer scientist can develop its own computer scripts to automate this process, a traditional biologist can find it difficult to stay up-to-date with the latest bioinformatics algorithms and software. We report here the creation of Armadillo, a software framework with a user-friendly interface developed in Java, which can be used for performing common tasks of phylogenetic analysis, including data-mining algorithms, multiple sequence alignments (MSA) and inferring of phylogenetic trees and networks. For instance, our software allows researchers to visualize the differences in the alignments obtained with different MSA algorithms (e.g., ClustalW, Muscle, Kalign), while exploring the effect of different evolutionary models on the tree building. At the same time, Armadillo can be used for the creation of data collections, methodologies, and complex computational pipelines for performing phylogenetic analysis.

Ph.D. 10 : Norman J. MacDonald :: Dalhousie University, Halifax, Nova Scotia

MITIGATING THE EFFECT OF SHARED ANCESTRY IN GENOTYPE-PHENOTYPE ASSOCIATION PROBLEMS WITH CONDITIONAL MUTUAL INFORMATION

(see oral presentation abstract)

Ph.D. 11 : Adriana Munoz :: University of Ottawa, Ottawa, Ontario

REARRANGEMENT PHYLOGENY OF GENOMES IN CONTIG FORM

(see oral presentation abstract)

Ph.D. 12 : Marie Pier Scott-Boyer :: Université de Montréal, Montréal, Québec

MODELING GENE EXPRESSION NETWORKS WITH GENOMIC DATA TO IDENTIFY DETERMINANTS OF COMPLEX QUANTITATIVE TRAITS IN MOUSE HEARTS

Complex traits, such as left ventricular mass, are phenotypic characteristics that result from interactions between a great number of genes and environmental factors. We propose to integrate gene expression networks with genomic and biologic data in order to refine the methods to construct gene expression networks and use that information to identify genetic variants that contribute to quantitative complex traits. We have use a panel of 24 mouse recombinant inbred strains (RIS) originating from crosses between A/J (A) and C56BL/6J (B) mice that we had studied previously to identify quantitative trait locus (QTLs) linked to the size of hearts. On the basis of whole genome microarrays (Illumina Mouseref-8 expression beadchips), we have enhanced the information of this dataset obtaining the transcriptional profile of ~20,000 non-redundant annotated genes in the hearts of all 24 AxB/BxA RIS. We will integrate these data with other types of information that include the genomic framework, the strength of correlation of gene expression with phenotypes of interest, the identity of genes influenced by expression QTLs, the relative positions of genes within loci and the existence of linkage disequilibrium blocks. We will use computational approaches to test the combined influences of these variables on the construction of gene expression networks.

Ph.D. 13 : Karine St-Onge :: Université de Montréal, Montréal, Québec

COMPARATIVE ANALYSIS OF ALLOSTERIC PATTERNS OF MIR-125A SINGLE BASE PAIR MUTANTS PREDICTS PROCESSING ACTIVITY OF THEIR PRIMARY TRANSCRIPTS

A single nucleotide polymorphism (SNP) of miR-125a is associated with breast cancer. A GC base pair in the major allele is mutated in a UC mismatch in the minor allele. This SNP influences its processing by the DROSHA/DGCR8 complex. We measured the processing activity of the 16 base pairs at the SNP position. Here, we present how using the MC-Fold secondary structure prediction program to generate the allosteric patterns of the miR-125a mutants predicts the effect of each mutant on its processing activity. We trained a support vector machine with integer vectors representing the allosteric patterns of the 16 variants. The predicted activities of each variants correlate (r^2 = 0.62) with the measured processing values by RT-QPCR. Since this approach is based solely on single sequence information, knowing the wild-type miRNA also predicts the possible effect on the function of variants and suggests it can be used to study SNPs in other non-coding RNAs as well.

Post-doc. 1 : Kelil Abdellali :: Université de Montréal, Montréal, Québec

ALIGNER: A NOVEL ALGORITHM FOR DETECTING AND ALIGNING RELATED PROTEIN SEQUENCES

(See oral presentation abstract)

Post-doc. 2 : Greg Finak :: Université de Montréal, Montréal, Québec

MERGING MIXTURECOMPONENTS FOR CELL POPULATION IDENTIFICATION IN FLOW CYTOMETRY

High throughput flow cytometry (HTFCM) is a relatively novel area utilizing established techonology (flow cytometry) for high-content screening and diagnostics. It is characterized by high dimensional data sets comprising hundreds to thousands of samples. A key step in the analysis of FCM data is gating, the identification of cell populations of interest (usually performed manually) from a series of bivariate dot plots. This traditional paradigm of data analysis can not keep up with the pace of the technology. It rapidly becomes inefficient in a highly multidimensional setting. A number of strategies have been proposed to perform automated gating based on different forms of mixture models. However, the complexity of cell populations in flow data has required users to compromise between model fit and an accurate estimate of the true number of cell populations.We present a framework for automating the gating of cell sub–populations from flow cytometry data based on merging mixture components using the flowClust methodology. We show that our cluster merging algorithm under our framework improves model fit and provides a better estimate of the number of distinct cell sub–populations than either gaussian mixture models or flowClust, especially for complicated flow cytometry data distributions. Our framework allows the automated selection of the number of distinct cell sub–populations and we are able to identify cases where the algorithm fails, thus making it suitable for application in a high throughput FCM analysis pipeline. Furthermore, we demonstrate a method for summarizing complex merged cell sub–populations in a simple manner that integrates with the existing flowClust framework and enables downstream data analysis. We demonstrate the performance of our framework on simulated and real FCM data. The software is available in the flowMerge package through the Bioconductor project.

Post-doc. 3 : Robert Flight :: Dalhousie University, Halifax, Nova Scotia

NORMALIZATION METHODS FOR TIME-COURSE DNA MICROARRAY DATA

(See oral presentation abstract)

Post-doc. 4 : Emmanuel Levy :: Université de Montréal, Montréal, Québec

ARE ALL PROTEIN-PROTEIN INTERACTIONS FUNCTIONAL? LESSONS FROM EVOLUTION

(See oral presentation abstract)

Post-doc. 5 : Chérif Mballo :: Université du Québec à Montréal, Montréal, Québec

DATA MINING TECHNIQUES FOR THE PREDICTION OF EXPERIMENTAL HIGH THROUGHPUT SCREENING (HTS) DATA

High Throughput Screening (HTS) remains a very costly process notwithstanding many technological advances in the field of biotechnology. Data mining has been described as "the exploration and analysis, by automatic or semi-automatic means, of large quantities of data to discover meaningful patterns and rules" (Berry and Linoff, 1997). Some statistical methods have been recently proposed to address the needs of experimental HTS campaigns (Brideau et al. 2003, Makarenkov et al. 2007). In this paper, we discuss the possibility of using data mining techniques to predict experimental HTS measurements and thus reduce the cost of an HTS campaign. Such a virtual HTS analysis is based on the results of real HTS campaigns carried out with similar compound libraries and similar drug targets. In this way, we analyze an experimental HTS assay from the McMaster University Data mining and docking competition (Elowe et al., 2005) using binary decision trees, neural networks, support vector machines, linear discriminant analysis, partial least squares and k-nearest neighbors methods. First, we study separately the molecular and atomic descriptors in order to establish which of them provide a better discrimination. We use a stepwise variable selection to obtain a combined data set containing both molecular and atomic descriptors. We present and discuss the results provided by the considered machine learning methods in terms of identification of false positive and false negative hits in the combined data set.

Post-doc. 6 : Nicolas Rodrigue :: University of Ottawa, Ottawa, Ontario

MUTATION-SELECTION MODELS OF CODING SEQUENCE EVOLUTION WITH SITE-HETEROGENEOUS AMINO ACID FITNESS PROFILES

Modeling the interplay between mutation and selection at the molecular level is key to molecular evolutionary studies. To this end, codon-based evolutionary models have been proposed as pertinent means of studying long-range evolutionary patterns, and are widely used. However, these approaches have not yet consolidated results from amino acid level studies showing that purifying selection in proteins displays strong site-specific effects, which translate into heterogeneous amino acid propensities across the columns of alignments; related codon-level studies have instead focused on either modeling a single selective context for all codon columns, or a separate selective context for each codon column, with the former strategy deemed too simplistic and the latter deemed over-parameterized. Here, we integrate recent developments in nonparametric statistical approaches to propose a probabilistic model that accounts for the heterogeneity of amino acid fitness profiles across the coding positions of a gene. We apply the model to a dozen real protein-coding gene alignments and find it to produce biologically plausible inferences, for instance, as pertaining to site-specific amino acid constraints, as well as distributions of scaled selection coefficients. In their account of mutational features as well as the heterogeneous regimes of purifying selection at the amino acid level, the modeling approaches studied here can form a backdrop for several extensions, accounting for other selective features, for variable population size, or for subtleties of mutational features, all with parameterizations couched within population-genetic theory.

Participants

Nom	Institution	Courriel
Adbellali, Kelil	UdeM	aabdellali.kelil@umontreal.ca
Badescu, Dunarel	UQAM	badescu.dunarel@courrier.uqam.ca
Bemmo, Amandine	U. McGill	amandine.bemmo@mail.mcgill.ca
Blanchet, Marc-Frédérick	UdeM	marc.frederic.blanchet@umontreal.ca
Blanchette, Mathieu	U. McGill	blanchem@mcb.mcgill.ca
Boc, Alix	UQAM	alix.boc@uqam.ca
Boisvert, Sébastien	U. Laval	Sebastien.Boisvert.3@ulaval.ca
Burger, Gertraud	U.Montréal	Gertraud.burger@umontreal.ca
Eveleigh, Robert	U. Dalhousie	rb821606@dal.ca
Finak, Greg	IRCM	greg.finak@ircm.qc.ca
Flight, Robert	U. Dalhousie	rflight79@gmail.com
Fourati, Slim	UdeM	slim.fourati@umontreal.ca
Godzaridis, Elenie	U. Laval	elenie.godzaridis.1@ulaval.ca
Gottardo, Raphael	IRCM	Raphael.Gottardo@ircm.qc.ca
Grenier, Jean-Christophe	UdeM	jean-christophe.grenier@umontreal.ca
Hickey, Glenn	U. McGill	hickey@mcb.mcgill.ca
Hou, Siyuan	U. Dalhousie	s.hou@dal.ca
Hussin, Julie	UdeM	Julie.hussin@umontreal.ca
Kim, Ethan	U. McGill	ethan@cs.mcgill.ca
Kleinman, Claudia Laura	UdeM	cl.kleinman@umontreal.ca
Lang, B. Franz	UdeM	Franz.lang@umontreal.ca
Lefebvre, François	UdeM	Francois.Lefebvre.2@umontreal.ca
Lemieux-Perreault, L.-Philippe	UdeM	louis-philippe.lemieux.perreault@umontreal.ca
Levy, Emmanuel	UdeM	emmanuel.levy@gmail.com
Lord, Étienne	UQAM	lord.etienne@courrier.uqam.ca
Luo, Xuemei	U. Carleton	xmluo@hotmail.com
MacDonald, Norman J	U. Dalhousie	norman@cs.dal.ca
Major, François	UdeM	Francois.major@umontreal.ca
Mballo, Chérif	UQAM	mballo.cherif@courrier.uqam.ca
Medjahed, Seyyid Ahmed	U. Mostaganem	inf_medjahed@yahoo.fr
Mercier, Éloi	UdeM	Eloi.Mercier@ircm.qc.ca
Munoz, Adriana	U. Ottawa	amuno010@uottawa.ca
Parida, Laxmi	Watson/IBM	parida@us.ibm.com
Poujol, Raphael	UdeM	raphael.poujol@umontreal.ca
Quackenbush, John	Harvard U.	johnq@jimmy.harvard.edu
Reatha, Sandie	U. Ottawa	rsand102@uottawa.ca
Robichaud, Marie	U.Montréal	Marie.robichaud@umontreal.ca

Rodrigue, Nicolas	U. Ottawa	nicolas.rodrigue@uottawa.ca
Roure, Béatrice	UdeM	beatrice.roure@umontreal.ca
Scott-Boyer, Marie Pier	UdeM	marie.pier.scott-boyer@umontreal.ca
Segal, Eran	Weizmann Inst.	eran.segal@weizmann.ac.il
Shateri Najafabadi, Hamed	U. McGill	hamed.shaterinajafabadi@mail.mcgill.ca
Stormo, Garry	Washington U.	stormo@wustl.edu
St-Onge, Karine	UdeM	chatonn@hotmail.com
Wong, Dennis	U. Dalhousie	dennis.wong@dal.ca