COLLOQUE BIO-INFORMATIQUE
ROBERT CEDERGREN
BIOINFORMATICS COLLOQUIUM

Programme

Université de Montréal
3-4 novembre 2008

Présentations orales et affiches
Poster and oral presentations

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 5e colloque bio-informatique Robert-Cedergren !

Ce colloque 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'un concours d'affiches et de présentations orales et de rendre compte de l'importance grandissante de la bio-informatique dans les sciences de la vie.

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

- **Richard Bonneau,** Biology & Computer Science, New York University, USA
- **Mark Gerstein,** Molecular Biophysics & Biochemistry Department, Yale University, USA
- **Daniel Gautheret**, Institut de génétique et microbiologie, Université Paris-Sud, France
- **Lior Pachter**, Mathematics & Computer Science, University of California at Berkeley, USA

Au total, quinze présentations orales et treize affiches seront en lice dans cinq catégories.

Les prix individuels seront décernés dans les catégories suivantes :

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

Un excellent colloque bio-informatique à tous et à toutes !

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

Welcome to the 5th annual Robert Cedergren Bioinformatics Colloquium!

This fifth Colloquium is an annual event gathering the universitary community working in Bioinformatics. The main purpose of this event is to share the latest Bioinformatics developments, by posters and oral presentations to take into account the increasing role of Bioinformatics in life sciences.

Keynote speakers will be;

- **Richard Bonneau,** Biology & Computer Science, New York University, USA
- **Mark Gerstein,** Molecular Biophysics & Biochemistry Department, Yale University, USA
- **Daniel Gautheret**, Institut de génétique et microbiologie, Université Paris-Sud, France
- **Lior Pachter**, Mathematics & Computer Science, University of California at Berkeley, USA

This year, 15 oral presentations and 13 posters will compete in 5 categories.

Individual awards will be given in the following categories:

	Best oral presentations	Best posters
MSc	$ 1000	$ 500
Ph.D.	$ 1000	$ 500
Post-doc.	$ 1000	n/a

Enjoy this 5th annual Robert-Cedergren Bioinformatics Colloquium!

Gertraud Burger, Ph.D.
Leader
Bioinformatics graduate programs

Comités/Committees

Comité d'organisation / Organizing Committee

Gertraud Burger
Marie Robichaud
Elaine Meunier
Philipe Lampron

Juges / Referees

Gertraud Burger (UdeM)
Serguei Chteinberg (UdeM)
Miklós Csurös (UdeM)
Nadia El-Mabrouk (UdeM)
Sylvie Hamel (UdeM)
Damian Labuda (CHU Ste-Justine)
Franz Lang (UdeM)
Nicolas Lartillot (UdeM)
François Major (IRIC)
Vladimir Makarenkov (UQAM)
Stephen Michnick
Hervé Philippe (UdeM)
Reza Salavati (McGill)
Marcel Turcotte (U. Ottawa)

Modérateurs / Session Chairs

Gertraud Burger
Franz Lang
Marcel Turcotte
Sylvie Hamel

Renseignements généraux / General information

Accueil / Registration

L'accueil des participants se fera au Hall d'honneur du pavillon Roger-Gaudry (anciennement Pavillon principal), le lundi 3 novembre dès 8 h 30. Les insignes d'identification vous seront remis à ce moment.

Les présentations orales auront lieu dans la salle M-415, alors que les affiches seront exposées dans le Hall d'Honneur.

The registration office is located in the Honor Hall in the Roger-Gaudry Building (previously Main Building). Your identification badge will be available from 8:30 am, November 3.

Oral presentation will take place in Room M-415 and poster session will take place in the Honour Hall.

Pauses santé et cocktail / Coffee breaks and cocktail

Les pauses santé et le lunch seront servis dans le Hall d'honneur.

Coffee breaks and the lunch will be served in the Honor Hall.

BIOINFORMATIC ROBERT CEDERGREN COLLOQUIUM 2008

November 3			
	Time	**Oral Presentations (room M-415)**	**Poster Presentations (Honor Hall)**
Session 1	9:15	Colloquium opening : **Gertraud Burger**, Professor, Biochemistry department, UdeM, Colloquium organizer (M-415)	
	9:30	Conference 1 : **Richard Bonneau**, Biology & Computer Science, New York University, USA *Learning Dynamic Regulatory Networks*	
	10:30	Coffee break	
	11:00	Presentation M.Sc. 1 : Mathieu Lavallée-Adam :: McGill University *Detection of locally over-represented GO terms in protein-protein interaction networks*	
	11:30	Presentation M.Sc. 2 : Julie Hussin :: Université de Montréal *Natural selection eliminates detrimental and favors advantageous phenotypes*	
	12:00	Presentation M.Sc. 3 : Sébastien Boisvert :: Université Laval *The distant segments kernel and the support vector machine : an alignment-free method for HIV type 1 coreceptor usage prediction*	
12:30		Lunch + poster session	
Session 2	13:30	Presentation M.Sc. 4 : Sandie Reatha :: University of Ottawa *Predicting the emergence of Influenza A viruses*	
	14:00	Presentation M.Sc. 5 : Brady Tracey :: University of Ottawa *Are Host- or Subtype-Switches Associated with Adaptive Evolution in Influenza A?*	
	14:30	Presentation Ph.D. 1: Yan Zhou :: Université de Montréal *Mixture Covarion model*	
	15:00	Coffee break + poster session	
	15:30	Presentation Ph.D. 2 : Slim Fourati :: Université de Montréal *Gene-expression signatures predicting the response to tamoxifen in breast cancer patients*	
	16:00	Presentation Ph.D. 3 : Malika Aid :: Université de Montréal *DNA motif discovery approach adapted to ChIP-chip data*	
	16:30	Conference 2 : **Mark Gerstein**, Molecular Biophysics & Biochemistry Department Yale University, USA *Human Genome Annotation*	

Session 1	Session Chair : Gertraud Burger Referees : Nicolas Lartillot, Damian Labuda, Gertraud Burger
Session 2	Session Chair : Franz Lang Referees : Nicolas Lartillot, Serguei Chteinberg, Franz Lang
Posters referee	Miklós Csurös, Nadia El-Mabrouk, Stephen Michnick

BIOINFORMATIC ROBERT CEDERGREN COLLOQUIUM 2008

November 4			
Time		**Oral Presentations (room M-415)**	**Poster presentations (Honor Hall)**
Session 3	9:00	Conférence 3 : **Daniel Gautheret**, Institut de génétique et microbiologie, Université Paris-Sud, France *One shot Classification of all RNAs from a Bacterial Genome with Phylogenetic Profiling* (M-415)	
	10:00	Presentation Ph.D. 4 : Mathieu Lajoie :: Université de Montréal *Inferring evolutionary history of tandemly arrayed genes*	
	10:30	Coffee break (Honor Hall)	
	11:00	Presentation Ph.D. 5 : Yaoqing Shen :: Université de Montréal *The convoluted evolution of acyl-CoA dehydrogenase family*	
	11:30	Presentation Ph.D. 6 : Hamed Shateri Najafabadi :: McGill University *Codon usage as a universal mechanism for synchronization of gene expression*	
	12:00	Presentation Post-doc. 1 : Sivakumar Kannan :: Université de Montréal *Automated (RNA) motifs discovery in the mitochondrial genome of Diplonema papillatum*	
12:30		Lunch + poster session (Honor Hall)	
Session 4	13:30	Presentation Post-doc. 2 : Simon Joly :: McGill University *A novel method for distinguishing hybridization from incomplete lineage sorting*	
	14:00	Presentation Post-doc. 3 : Ali Mokdad :: Université de Montréal *Three-dimensional modeling of human precursor MicroRNAs*	
	14:30	Presentation Post-doc 4 : Nicolas Rodrigue :: University of Ottawa *Phenomenological modeling of site-heterogeneities in protein-coding nucleotide sequence evolution using the Dirichlet process prior*	
	15:00	Coffee break + poster session (Honor Hall)	
	15:30	Conference 4 : **Lior Pachter**, Mathematics and Computer Science, University of California at Berkeley, USA *Finding the trees in Darwin's forest* (M-415)	
16 :30		Awards : **Muriel Aubry,** Biochemistry Dept. & **Jean Meunier**, Head of Computer Science Dept. (M-415)	
16 :50		Colloquium closing : **Gertraud Burger**, Organizer (UdeM) (M-415)	
17:00		Cocktail (Honor Hall)	

Session 3	Session chair : Marcel Turcotte Referees : Hervé Philippe, François Major, Marcel Turcotte
Session 4	Session chair : Sylvie Hamel Referees : Hervé Philippe, François Major
Posters referee	Nadia El-Mabrouk, Vladimir Makarenkov, Reza Salavati

CONFERENCES

Richard Bonneau, Biology & Computer Science, New York University, USA

Learning Dynamic Regulatory Networks

Learning regulatory networks from genomics data is one of the most important problems in biology today, with applications spanning biology and biomedicine. There are, however, a lot of reasons to believe that regulatory network inference is beyond our current reach due to the combinatorics of the problem, factors we can't (or don't often) collect genome wide measurements for, and dynamics that elude cost-effective experimental designs. In spite of these challenges multiple groups have recently shown that we can reconstruct large fractions of many prokaryotic regulatory networks from compendiums of genomics data and that these global regulatory models can be used to predict the dynamics of the transcriptome. These global regulatory models can be combined with modeling of metabolic and signaling networks to model the global operation of cells with unprecedented completeness and accuracy. We review an overall strategy for the reconstruction of global networks resulting from the recent progress of several genomics consortia.

Mark Gerstein, Molecular Biophysics & Biochemistry Department, Yale University, USA

Human Genome Annotation

A central problem for 21st century science will be the annotation and understanding of the human genome. My talk will be concerned with topics within this area, in particular annotating pseudogenes (protein fossils), binding sites, CNVs, and novel transcribed regions in the genome. Much of this work has been carried out in the framework of the ENCODE and modENCODE projects. In particular, I will discuss how we identify regulatory regions and novel, non-genic transcribed regions in the genome based on processing of tiling array and next-generation sequencing experiments. I will further discuss how we cluster together groups of binding sites and novel transcribed regions. Next, I will discuss a comprehensive pseudogene identification pipeline and storage database we have built. This has enabled us to identify >10K pseudogenes in the human and mouse genomes and analyze their distribution with respect to age, protein family, and chromosomal location. I will try to interrelate our studies on pseudogenes with those on transcribed regions. At the end I will bring these together, trying to assess the transcriptional activity of pseudogenes. Throughout I will try to introduce some of the computational algorithms and approaches that are required for genome annotation -- e.g., the construction of annotation pipelines, developing algorithms for optimal tiling, and refining approaches for scoring microarrays.

Daniel Gautheret, Institut de génétique et microbiologie, Université Paris-Sud, France

One shot Classification of all RNAs from a Bacterial Genome with Phylogenetic Profiling

Identification and characterization of functional elements in the non-coding regions of genomes is an elusive and time consuming activity whose output does not keep up with the pace of genome sequencing. Hundreds of bacterial genomes lay unexploited in terms of non-coding sequence analysis although they may conceal a wide diversity of novel RNA genes, riboswitches or other regulatory elements. We describe a strategy that exploits the entirety of available bacterial genomes to rapidly classify conserved non-coding elements from any reference species. This method clusters non-coding elements based on their profile of presence among species. Most non-coding RNAs display specific signatures that enable their efficient classification in distinct clusters, away from sequence conservation noise and other elements such as promoters. Besides offering a powerful method for de novo ncRNA identification, the analysis of phylogenetic profiles opens a new path towards the identification of functional relationships between co-evolving coding and non-coding elements.

Lior Pachter, Mathematics and Computer Science, University of California at Berkeley, USA

Finding the trees in Darwin's forest

The problem of determining homology among multiple related biological sequences, known as the alignment problem, is arguably the fundamental problem in comparative genomics. Accurate alignment is essential for both functional and evolutionary genomics studies. We explain how the problem of determining homology at the nucleotide level can be interpreted as finding the trees in "Darwin's forest" and focus on the tractability of the problem. We argue that many recent negative results emphasizing uncertainty in alignment are misleading in that they confound uncertainty in the choice of model, uncertainty in alignment given a model, and error due to heuristics used for inference. We explain how hidden Markov models for pairwise alignment can be extended to provide effective models for multiple alignment, and show that these models indicate little uncertainty in alignment of both unrelated sequences and of orthologous sequences from related species. Moreover, we discuss an algorithm that provides an efficient approach to finding the alignments with highest expected accuracy. Together, these results provide a path to the removal of lingering doubts about the accuracy of multiple alignments.

ORAL PRESENTATIONS

Oral M.Sc.

M.Sc. 1 : Mathieu Lavallée-Adam :: McGill University

Detection of locally over-represented GO terms in protein-protein interaction networks

High-throughput methods for identifying protein-protein interactions produce increasingly complex and intricate interaction networks. These networks are extremely rich in information, but extracting biologically meaningful hypotheses from them and representing them in a human-readable manner is challenging. We propose a method to identify Gene Ontology terms that are locally over-represented in a sub-network of a given biological network. Specifically, we propose two methods to evaluate the degree of clustering of proteins associated to a particular GO term and describe four efficient methods to estimate the statistical significance of the observed clustering. We show, using Monte Carlo simulations, that our best approximation methods accurately estimate the true p-value, for random scale-free graphs as well as for actual yeast and human networks. When applied to these two biological networks, our approach recovers many known complexes and pathways, but also suggests potential functions for many sub-networks.

M.Sc. 2 : Julie Hussin :: Université de Montréal

Haplotype allelic classes and positive selection in the human genome

Natural selection eliminates detrimental and favors advantageous phenotypes. This process leaves characteristic signatures in the underlying genomic segments that can be recognized through deviations in the allelic or in haplotypic frequency spectra, analyzed usually within the frameworks of the infinitely many sites or the infinitely many alleles model, respectively. We introduce a new way of looking at the genomic single nucleotide polymorphisms : the haplotype allelic classes (HACs). The model combine segregating sites and haplotypic informations in order to reveal useful characteristics of the data, providing an identifiable signature of natural selection that can be identified by comparison with the background distribution. We compare the HACs distribution's partition between the haplotypes carrying the selected allele and the remaining ones. Coalescence simulations are used to study the distributions under standard population models assuming neutrality, demographic scenarios and selection models. To test, in practice, the performance of HACs and the derived statistics in capturing deviation from neutrality due to selection, we analyzed the genetic variation in the locus of lactase persistence in the three HapMap populations : european-derived population (USA), asian population (China and Japan) and african population (Yoruba from Nigeria). As expected, we found a strong signal of positive natural selection in the lactase persistence locus in the european-derived population whereas no signal has been identified in the other populations. Furthermore, we developed a method to scan the whole genome with the new HACs-based approach to identify other regions under positive selection and to compare our results with those obtained in other recent studies seeking for target of natural selection in the human genome.

M.Sc. 3 : Sébastien Boisvert :: Université Laval

The distant segments kernel and the support vector machine: an alignment-free method for HIV type 1 coreceptor USAGE PREDICTION

HIV type 1 infects human cells through the interactions between ligands and receptors. Accordingly, this retrovirus uses the CD4 receptor in conjunction with a chemokine receptor, to penetrate target cells. In vivo, the chemokine receptor is either CCR5 or CXCR4. Bioinformatic methods were described to predict the coreceptor usage but they all rely on sequence alignments, making any sequences with too many indels not processable. To cope with this drawback, we developped an alignment-free approach using string kernels and support vector machines. The SVM has strong theoretical support and is very robust to noise. We created a new string kernel, namely the distant segments kernel, and compared it to existing string kernels in the litterature, such as the local alignment kernel and the blended spectrum kernel.

We obtained, with the distant segments kernel, an accuracy (1-empirical risk) of 94.80% on a testing set of 1425 examples with a classifier trained on a set of 1425 examples. Our algorithm outperforms the current state-of-the-art method for this classification task. Out of the 1425 training examples, only 577 were used as support vectors by the support vector machine, which indicates that a large margin linear classifier exists in a large feature space. Our method allows the fast and accurate prediction of all allowed coreceptor usages, that are CCR5, CXCR4 and CCR5-and-CXCR4. We implemented a web server to perform automatic classification through the CGI interface. This web server is available at http://genome.ulaval.ca/hiv-dskernel.

Support vector machines and string kernels have broad applicability in bioinformatics, such as remote protein homology detection, gene finding, and clustering. Furthermore, kernels are not limited to bioinformatics, but can also be applied to many tasks in chemoinformatics, such as virtual screening trials.

M.Sc. 4 : Sandie Reatha :: University of Ottawa

Predicting the emergence of Influenza A viruses

Vaccine design for influenza viruses is based on an empirical guess of which influenza strains will most likely be circulating in the upcoming flu season. The choice of the strains to be included in the vaccine is crucial, as an erroneous decision can lead to an epidemic or even a pandemic. Here we present a prediction model of future influenza strains based on a realistic model that incorporates natural features of the virus such as selection, recombination and reassortment.

To simplify the computational burden, we use a sampling technique that permits us to use the extensive Influenza Virus Resource database by clustering sequences based on their pairwise similarity and eliminating sequences with 97% or higher similarity. The predictive power of our model is demonstrated with the case of the unexpected emergence of a new H3N2 strain in the 2007-08 flu season.

M.Sc. 5 : Brady Tracey :: University of Ottawa

Are Host- or Subtype-Switches Associated with Adaptive Evolution in Influenza A?

Past human Influenza A pandemics in the 20th were caused by viruses that switched from non-human hosts such as wild-waterfowl. Among the three major pandemics documented in the last century, Spanish Influenza (1918), Asian Influenza (1957) and Hong Kong Influenza (1968), reassortment is though to have played a key role in at least two of them, those of 1957 and of 1968. While both reassortment and recombination are known to play a role in the emergence of pandemics, it is unknown whether such changes of host are adaptive for the virus. Here we focus on Southeast Asia and investigate the adaptive nature of host switches for Influenza A viral complete genomes. We hypothesize that three processes drive the emergence of new subtypes that could lead to future pandemics: reassortment, host switch and positive selection. We dissect the interplay between these three processes by the analysis of complete Influenza A genomes of both human and avian hosts under three scenarios: (i) a non-epidemic season (2001-2002), (ii) a season with a mild outbreak (1996-1997) and (iii) the current H5N1 situation (2006-2008), whose pandemic potential is unclear.

We show that there is no clear association between adaptive evolution, host switch and the emergence of a pandemic. This suggests that pandemics result mostly from nonadaptive events.

Oral Ph.D.

Ph.D. 1 : Yan Zhou :: Université de Montréal

Mixture Covarion Model

Recently heterotachy (Lopez et al., 1999), in which scenario substitution rates vary not only across sites but also across time, has drawn many researchers' attention (Kolaczkowski and Thornton, 2004; Lockhart et al., 1996; Zhou et al., 2007). It has been shown that heterotachy widely exists in real datasets (Lopez et al., 1999) and potentially impedes the phylogeny inference (Kolaczkowski and Thornton, 2004; Lockhart et al., 1996). One way to handling heterotachy is the Covarion model (Huelsenbeck, 2002; Tuffley and Steel, 1998), in which there are two states: "on" and "off". In ON states, sites are available to be substituted; in OFF states, sites are not allowed to be substituted; the switch rates between ON and OFF are stationary across sites and along the branch. We show that the switch rates between ON and OFF vary across sites. Thus we develop an infinite mixture model to handle these heterogeneities of Covarion parameters across sites with Dirichlet process (Neal, 2000). Posterior predictive discrepancy tests (Gelman et al., 1996) show covarion mixture model has a better model fit than the non-mixture covarion model.

Ph.D. 2 : Slim Fourati :: Université de Montréal

Gene-expression signatures predicting the response to tamoxifen in breast cancer patients

Introduction: In Canada, one in nine women is diagnosed with breast cancer during the course of her life. 2/3 of mammary tumours expresse the estrogen receptors (ER) and are stimulated by estrogens. Treatment with antiestrogens such as tamoxifen can reduce tumor proliferation. Nevertheless, tamoxifen is efficient only in 50% of ER+ breast cancer cases. The histological and clinical factors used for therapeutic indication are insufficient to reflect the disease heterogeneity and to predict the success of antiestrogen adjuvant therapy. Several studies showed that estrogen-induced genes could act as prognostic markers (indicative of the rate of relapse) for ER+ breast cancer patients. However the predictive value of these markers as indicators of the success of antiestrogen treatment remains to be examined.

Hypothesis: Our hypothesis is that estrogen-induced genes can also serve as predictive markers of tamoxifen response for ER+ breast cancer patients.

Results: DNA microarray experiments performed with MCF-7 human mammary tumor cells in Dr. Mader's laboratory identified 170 primary estrogen target genes. As a first step, we validated the prognostic value of estrogen target genes on a dataset of tumours treated with tamoxifen. To assess the predictive value of these genes (their capacity to identify a group of tumours that would benefit from the treatment) we used another dataset of tumours consisting of tamoxifen-treated and non-treated tumours. This dataset was divided into two subsets, one of which was used as training dataset (109 treated and 61 non-treated tumours). A non-supervised approach, k-means partitioning, was used to separate tumours treated by tamoxifen in two groups on the basis of the expression levels of primary target genes. A Kaplan-Meier analysis allowed us to determine that the two groups had significantly different rates of relapse. On the other hand, classification of the 61 tumours which received no adjuvant treatment in these groups (nearest centro•d prediction) does not allow to predict significantly different rates of relapse. These observations were validated on the test subset (159 treated patients and 64 patients non-treated by tamoxifen). Therefore, estrogen primary target genes can specifically identify the patients benefiting from tamoxifen therapy. The group that does not benefit from the treatment may require a more aggressive treatment. On the other hand, secondary target genes are prognostic of relapse and this independently of the treatment given to the patients.

Conclusion: Our study suggests that it is possible to predict the success of antiestrogen therapy through the expression levels of primary estrogen target genes, which could help refine therapeutic indication and improve survival of patients.

Ph.D. 3 : Malika Aid :: Université de Montréal

DNA motif discovery approach adapted to ChIP-chip data

Transcription factors play an important role in various biological processes such as differentiation, cell cycle progression and tumorigenesis. They regulate gene transcription by binding to specific DNA sequences (cis-regulatory elements). Identifying these cis-regulatory elements is a crucial step in the understanding of gene regulatory networks. The recent developments in genomic technologies such as DNA microarrays and Chromatin immuno-precipitation followed by microarray hybridization (ChIP-chip) have made possible the whole genome characterization of TFs binding sites (TFBS) and allowed the development of several computational DNA motif discovery tools. Although these various tools are widely used and have led to the discovery of novel motifs, in practice none of them have proven to be efficient in control data sets due to a high rate of false positive and false negative predictions.

DNA motif discovery tools use different strategies to extract and represent the motif patterns: Enumerative or alignment-based approaches. Each of them uses a specific scoring function to evaluate motifs and report those having the highest scores compared to a reference data set. The main drawback of these scoring functions is the fact that some motifs occurring ubiquitously in the genome are scored very highly, despite not being real enriched (false positive predictions). Consequently, real enriched motifs with low scores are penalized (false negative predictions).

Analyses conducted on simulated and ChIP-chip data using different tools: Mdmodule (enumerative algorithm, MEME (Expectation and maximization algorithm), and MotifSampler (Gibbs algorithm), have shown that DNA motif discovery tool scoring functions do not represent the observed characteristics of TFBS in ChIP regions. For example they do not take into account that motifs representing real binding sites are more likely to reside near the center of the ChIP regions. Our results showed that these scoring functions are not adapted to the discovery of TFBS in ChIP-chip data.

We propose to implement two new scoring functions: motif positional bias and motif group specificity that take into consideration the characteristics of the distribution of TFBS in the ChIP regions. Motif positional bias measures how a given motif is distributed across the ChIP regions. It is expected that true binding sites will be enriched in specific positions (peak in the ChIP central regions) compared to what is expected by chance (background data set). Group specificity score, is a measure of how a given motif targets the set of Chip regions. We expect that true TFBS will be distributed evenly throughout the ChIP sequences and are clearly more frequent compared to a reference data set.

We applied these scoring functions on a simulated data set and on real ChIP data set. The results show that our approach enhances the DNA motif discovery tools predictions and significantly reduce the rate of false positive predictions.

Ph.D. 4 : Mathieu Lajoie :: Université de Montréal

Inferring evolutionary history of tandemly arrayed genes

Tandemly arrayed genes (TAGs) represent a large fraction of most genomes and are involved in many biological processes, such as recognition and signal transduction (olfactory receptors), regulation of gene expression (zinc-finger genes) and molecular transport (globins). TAGs evolve mainly through unequal crossing over during meiosis, which can duplicate one or more adjacent genes simultaneously (tandem duplication). Many algorithms have been proposed to infer a tandem duplication history for a TAG family. However, their applicability is often limited in practice, because other evolutionary events such as inversions and inverted duplications also contribute to TAG evolution. This is highlighted by the fact that many TAG clusters contain genes in both transcriptional orientations. Recently, we proposed the first inference algorithm which models both tandem duplications and inversions. Here we present a new algorithm which models tandem duplications, inverted duplications, gene losses and inversions.

Ph.D. 5 : Yaoqing Shen :: Université de Montréal

The convoluted evolution of acyl-CoA dehydrogenase family

Acyl-CoA dehydrogenases (ACAD) constitute a large and diverse protein family of at least 11 subclasses. Some subclasses participate in the first step of mitochondrial beta oxidation (i.e., the short-, medium-, long- and very long-chain acyl-CoA dehydrogenases, ACADS, ACADM, ACADL, ACADV, and ACAD9, respectively). Others are involved in amino acid degradation; these are iso(3)valeryl-CoA dehydrogenase (IVD), iso(2)valeryl-CoA dehydrogenase (ACDSB), isobutyryl-CoA dehydrogenase (ACAD8), and glutaryl-CoA dehydrogenase (GCAD). Two additional subclasses, ACAD10 and ACAD11, are found in human, but their function remains unknown. In many instances, proteins are annotated as ACAD without specifying to which subclass they belong. All three kingdoms of life possess ACAD genes, and the subclasses in eukaryotes were proposed to be obtained from alpha-proteobacteria through the endosymbiotic event that leads to mitochondria. Yet, evidence for this hypothesis is lacking. Therefore, a thorough phylogenetic study of this family is needed to understand the dispersal and relationship of two key metabolism pathways: beta oxidation of fatty acid, and amino acid degradation. Using complete genome sequences, we performed a large-scale screen for ACAD in 211 species including Archaea, Bacteria, and Eukaryotes, and we assigned specific function to sequences previously annotated as ACAD. Our phylogenetic analyses indicate horizontal gene transfer (HGT) of ACAD in particular within proteobacteria, which, in combination with multiple gene duplication events, highly complicate the inference of the evolution history. Nevertheless, we discovered several patterns: (1) Sequences from human and fungi (if they have the particular homolog) always group together, suggesting a common origin of this protein family in Opisthokonts. (2) Fungi lack several subclasses (e.g., ACAD8, ACAD9, ACADV, ACADS), but functionally, a single ACAD member combines the activity of ACAD8, ACADSB, and ACADS. (3) In fungi, ACADM occurs to only a few species (Basidiomycetes). Ascomycetes have instead homologs of fadE12, an enzyme of the same function, likely acquired from alpha-proteobacteria. (5) Subclasses involved in fatty acids degradation are absent from land plants, which is in agreement with the fact that plants do not have mitochondrial beta oxidation. In sum, the hypothesis of an alpha-proteobacterial origin of the ACAD family did not gain sufficient support. Nonetheless, we were able to elucidate the more recent family evolution history; resolving the ancient and apparently convoluted path will require more and taxonomically broader genome data.

Ph.D. 6 : Hamed Shateri Najafabadi :: McGill University

Codon usage as a universal mechanism for synchronization of gene expression

Understanding the biological mechanisms underlying the regulation of gene expression has been the subject of enormous studies, mostly focusing on the elements present in non-coding regions surrounding the ORFs of protein-coding genes. However, the nucleotide sequence of the coding region itself has been ignored vastly for its capability of regulating protein expression. In a recent study, we have shown that proteins that either physically or functionally interact with each other have similar synonymous codon usages. Here, we show that this observation is a consequence of similarity among synonymous codon usages of genes that are co-regulated. Using genome-wide expression profiles of four diverged organisms, human, yeast, Escherichia coli and Caenorhabditis elegans, we provide rigorous statistical analysis showing that genes with similar expression profiles have similar synonymous codon usages. Using functional groups as surrogates of clusters of coexpressed genes, we show that this pattern is universal and can be observed in almost all eukaryotes as well as bacteria and archaea. Furthermore, we show that this observation can neither be completely explained by average protein expression level nor by regional variations of genomic GC content. Finally, we propose a model in which codon usage is used to synchronize the changes in expression levels of co-expressed proteins, so that when the composition of tRNA pool is changed inside a cell, proteins that are related to each other respond similarly to such alterations.

Oral Post-Doctorat

Post-doc 1 : Sivakumar Kannan :: Université de Montréal

Automated (RNA) motifs discovery in the mitochondrial genome of Diplonema papillatum

In mitochondria of the unicellular eukaryote *Diplonema,* genes are systematically fragmented into small pieces that are encoded on separate chromosomes, transcribed individually, and then concatenated into contiguous messenger RNA molecules [1]. Similar to their sister group—Kinetoplastids, Diplonema mRNA also undergoes editing. It is hypothesized that guide RNAs (gRNAs) that direct the RNA editing are also involved in concatenating the fragmented transcripts. We have tested this hypothesis on the *cox1* gene of *Diplonema papillatum* that consists of nine pieces. A potential gRNA is a RNA segment with two elements each matching the boundary regions of two consecutive fragmented transcripts (*i*, *i*+1). The match length is expected to be between 6 and 10 nucleotides and the distance between the two elements should be less than 150 nucleotides. Several thousands of potential gRNAs were discovered matching the above-mentioned criteria and it is challenging to identify the most likely gRNAs. I will talk about the filtering procedure we developed to select the most likely candidates.

Post-doc 2 : Simon Joly :: McGill University

A novel method for distinguishing hybridization from incomplete lineage sorting

Hybridization is widespread in nature. Yet, the difficulty of distinguishing hybridization from incomplete lineage sorting makes the extent and evolutionary significance of hybridization difficult to evaluate. Although some methods have been described in the past to differentiate these two evolutionary processes, they do not have wide application. Here I present a novel method for statistically distinguishing hybridization from incomplete lineage sorting based on minimum genetic distances of a non recombining locus. It is based on the idea that the expected genetic distance between sequences from two species is smaller for some hybridization events than for incomplete lineage sorting scenarios. When applied on empirical datasets, distributions can be generated for the minimum inter-species distances expected under incomplete lineage sorting using coalescent simulations. If the observed distance between sequences from two species is smaller than its predicted distribution, incomplete lineage sorting can be rejected and hybridization inferred. I evaluate the power of the method using simulations for a range of parameter values. These include the number of individuals sampled per species, the population sizes, the time between speciation and the hybridization event and the time since the hybridization event, and the sequence length of the marker. Results show that the method has good power if the hybridization event is not to close to the speciation event and if sequences are of considerable length (> 1000 bp). I finally apply the method on an empirical dataset and identify an ancient hybridization event in the Late Tertiary radiation of the New Zealand alpine buttercups (Ranunculus). Because the method is robust and makes few assumptions, it is expected to be widely applicable and provides a rigorous approach to assess the importance of hybridization in evolution.

Post-doc 3 : Ali Mokdad :: Université de Montréal

Three-dimensional modeling of human precursor MicroRNAs

There are about 700 human precursor microRNA (pre-miRNA) entries in miRBase database(1), none of which with a known three-dimensional (3D) structure. It is not even conceivable to determine experimentally the structure of each of them. Our aim here is to determine a viable and accurate 3D model for each human pre-miRNA, and then to extract the 3D rules that allow these miRNAs to function in vivo. It is important to remember that at first glance pre-miRNAs seem like regular RNA helices with uninteresting 3D structural features, but in fact they are seeded with specifically positioned and oriented non-canonical interactions and bulges that can only be studied and characterized in 3D. Recently, our lab released a set of software tools that efficiently and accurately predict RNA 3D models from primary sequence data(2). With these tools we obtain a set of solutions that represent the 2D and 3D possibilities an RNA sequence can fold into. The problem is then reduced to identifying the best models from the pool of results. To achieve this we implement known structural constraints derived from the only crystal structure of Dicer with a modeled double stranded RNA helix docked with it(3). Based on this Dicer structure, the range of lengths of mature miRNA sequences, and the statistics of preliminary 2D and 3D structures of pre-miRNAs that we have already determined, we hypothesize that "human pre-miRNAs do not fold in an identical fashion to form a rigid stem with exact length, but rather fold according to several possible helical templates that range in size between 18 and 22 base or base pair steps between Drosha and Dicer scissile phosphates". Dicer is thought to be flexible enough to accommodate and cleave the range of these sizes(4). To identify the natural helical template for each pre-miRNA, we are currently folding each of them in 3D according to the distance constraints of each of five predetermined templates. We will then select the one template that produces most low free-energy structures. Pre-miRNAs that belong to each template will be structurally aligned, allowing for the determination of their endogenous structural design rules. An application of this project is to use the obtained 3D rational design rules to create therapeutic miRNAs with high efficiency and low toxicity compared to current artificial miRNAs(5).

Post-doc 4 : Nicolas Rodrigue :: University of Ottawa

Phenomenological modeling of site-heterogeneities in protein-coding nucleotide sequence evolution using the Dirichlet process prior

Using a nonparametric device known as the Dirichlet process prior, we propose Markovian models of codon substitution that recognize the heterogeneity of amino acid preferences across the coding positions of a gene. The basic model is constructed from a global specification of mutational properties, along with a mixture of amino acid preference profiles modulating selective constraints. Under the Dirichlet process, however, the exact form of the mixture is not predefined, with the number of components controlled by a higher level (hyper-) parameterization. A second (independent) Dirichlet process is also applied to further modulate nonsynonymous rate heterogeneity, without regard to the amino acid states involved; as such, two codon sites may be affiliated to the same amino acid preference profile, with one of these sites undergoing more changes between the allowed amino acids than the other. Bayesian assessments show the importance of such a model formulation in order to capture the basic features of protein-coding sequence evolution, and emphasize its usefulness in elucidating site-specific selective patterns operating at the amino acid level.

AFFICHES / POSTERS

Affiches M.Sc.

M.Sc. 1 : Diala Abd Rabbo :: Université de Montréal

Relation between transcryptome and genotype in normal tissues of the heterozygote females carrying a mutation on brca

Problem: Women carrying mutations on *BRCA1/2* are at high risk to develop ovarian cancers. Those cancers are often lethal since diagnosed at an advanced stage. There is an urgent need to improve diagnosis tools by identifying genes differentially expressed at an early stage of the hereditary carcinogenesis.

Hypothesis: An expression profile is associated to a mutation of *BRCA1/2* in non-tumor ovarian surface epithelium (NOSEs) cells.

Material: The expression profiles of the selected samples were produced with the Affymetrix microarray HuFL 6800®. RNA samples were extracted from 9 primary cultures of NOSEs : 4 non-carriers (class1), 2 carriers of a BRCA1 mutation (class 2) and 3 of a *BRCA2* mutation (class 3), and 3 ovarian tumors (TOVs) from carriers of a *BRCA1* mutation.

Methods: We performed a supervised differential analysis, using LIMMA (Linear Model for Microarrays Data) package from Bioconductor®, to identify the expression profiles associated to a *BRCA1/2* mutation. To confirm the accuracy of our microarray results, we used real-time quantitative PCR (q-RT-PCR) to quantify a set of 4 candidate genes.

Results: In this pilot study, we showed the existence of a molecular profile associated to the presence of a mutation in the NOSE cells. We validated by q-RT-PCR, the pattern of expression of 4 candidate genes that were at the top rank of the gene list produced by our LIMMA analysis.

Perspective: We want to confirm results of this pilot study on a larger series. Moreover, we aim to look for co-regulated chromosomal regions associated to a mutation of *BRCA1/2*.

M.Sc. 2 : Mhamedi Abdellah Alaoui :: Université de Montréal

Des arbres phylogénétiques pour la modélisation de la structure secondaire

La structure secondaire joue un rôle important dans le fonctionnement des cellules et elle a aussi un impact sur les applications thérapeutiques en médecine. Knudsen et Hein (1999) ont proposé des modèles simples basés sur des grammaires stochastiques pour sa prédiction. La vraisemblance d'une structure secondaire dans ces modèles peut s'écrire en termes des probabilités de transition des acides aminés associées à des arbres phylogénétiques qui décrivent leur évolution. Nous présentons des algorithmes pour le calcul des arbres phylogénétiques basés sur des groupements heuristiques des données. Notre critère de similarité est la vraisemblance associée aux différentes topologies particulières des branches de l'arbre impliquant plusieurs arrangements d'acides aminés à la fois. Nos probabilités de transition suivent le modèle de Muse (1994), une variante du modèle de Hasegawa-Kishino-Yano, qui est basé seulement sur trois paramètres pour décrire les transitions des seize paires d'acide aminés. Nous présentons aussi un modèle de Poisson qui permet l'estimation de ceux-ci.

M.Sc. 3 : Sébastien Boisvert :: Université Laval

The distant segments kernel and the support vector machine : an alignment-free method for HIV type 1 coreceptor usage prediction

HIV type 1 infects human cells through the interactions between ligands and receptors. Accordingly, this retrovirus uses the CD4 receptor in conjunction with a chemokine receptor, to penetrate target cells. In vivo, the chemokine receptor is either CCR5 or CXCR4. Bioinformatic methods were described to predict the coreceptor usage but they all rely on sequence alignments, making any sequences with too many indels not processable. To cope with this drawback, we developped an alignment-free approach using string kernels and support vector machines. The SVM has strong theoretical support and is very robust to noise. We created a new string kernel, namely the distant segments kernel, and compared it to existing string kernels in the litterature, such as the local alignment kernel and the blended spectrum kernel. We obtained, with the distant segments kernel, an accuracy (1-empirical risk) of 94.80% on a testing set of 1425 examples with a classifier trained on a set of 1425 examples. Our algorithm outperforms the current state-of-the-art method for this classification task. Out of the 1425 training examples, only 577 were used as support vectors by the support vector machine, which indicates that a large margin linear classifier exists in a large feature space. Our method allows the fast and accurate prediction of all allowed coreceptor usages, that are CCR5, CXCR4 and CCR5-and-CXCR4. We implemented a web server to perform automatic classification through the CGI interface. This web server is available at http://genome.ulaval.ca/hiv-dskernel. Support vector machines and string kernels have broad applicability in bioinformatics, such as remote protein homology detection, gene finding, and clustering. Furthermore, kernels are not limited to bioinformatics, but can also be applied to many tasks in chemoinformatics, such as virtual screening trials.

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

A comparison of microarray differential expression detection methods based on consistency with functional annotations

In the field of microarray data analysis, many preprocessing methods and statistical procedures are available to assess differential expression. Yet it is not possible to identify a gold standard analysis pipeline that is globally accepted by the scientific community, even within a specific microarray platform. This has serious implications since different pipelines often yield quite dissimilar lists of differentially expressed genes. In this work we hypothesize that the interdependency between gene expression levels, through the coordination of regulation within functional classes, can be used as a mean to quantitatively compare the above mentioned analysis pipelines. Since functionally related genes are likely to be simultaneously differentially expressed, the best pipeline should yield lists of differentially expressed genes more highly biased towards lists of genes already known from the literature to be functionally related. By measuring this bias across many microarray experiments and many such lists, we expect the best method to stand out. Experimental datasets from the Affymetrix platforms HGU133A and HGU133 Plus 2.0 were retrieved from the National Center for Biotechnology Information Gene Expression Omnibus, providing us with over 800 contrasts (group comparisons). For each contrast, gene rankings were obtained by following various differential expression analysis pipelines. Consistency between those rankings and gene sets collection from the MsigDB database [Subramanian 2005] was assessed using either the Mann-Whitney U (area under the ROC curve) or the Kolmogorov-Smirnov statistic. Results indicate the fold-change criterion to yield gene rankings significantly more consistent with the chosen functional annotations than other gene selection criterion. This somewhat surprising conclusion resonates with the controversial finding of the MAQC consortium [MAQC Consortium 2006] on the use of the fold-change to maximize reproducibility.

M.Sc. 5 : Yuan Mao :: McGill University

Putative structural RNAs in Trypanosoma brucei

Trypanosoma brucei is a parasitic organism causing African Sleeping Sickness among the Sub-Saharan African population. It belongs to a group of organisms, called trypanosomatids, which have diverged very early from other eukaryotes and possess a distinct system of genetic control that relies heavily on a variety of structural RNA elements. These RNA elements serve in regulating the expression of different genes by providing signals for a variety of post-transcriptional events, such as splicing, regulation of RNA stability, editing and translation. Computational identification of structural RNAs has been a tedious task due to their variability in sequence features and composition. Here, we present a genome-wide analysis for identification of structural RNAs in T. brucei. Using suitable positive and negative controls, we estimate a sensitivity of about 50% and a precision of about 90%. We obtained a set of 205 potential candidates whose sequence and structure are conserved between T. brucei and its diverged relative, Leishmania braziliensis. These include known and putative novel rRNAs, tRNAs, snRNAs and regulatory elements at 5' and 3' UTRs of protein-coding genes as well as novel structural RNAs that may not fit into any of the above categories.

M. Sc. 6 : Philippe Nadeau :: Université de Montréal

Testing haplotype allelic classes and associated statistics for positive selection detection with coalescence simulations

Mutations, present in a population or a species and not fixed, are called polymorphisms. Single Nucleotide Polymorphisms (SNPs), due to nucleotide substitutions, are usually bi-allelic. Contiguous SNPs on the same chromosome form a haplotype. Haplotypes present in a population sample can be organized into haplotype allelic classes (HACs) grouping haplotypes with the same mutational distance from the reference haplotype, such as, for example, the ancestral haplotype or the one that contain all major alleles. HACs combine the information on the frequency spectra of polymorphic sites and haplotypes, usually analyzed separately, within the framework of the infinitely-many-sites and the infinitely-many-alleles model, respectively. HACs distribution and occupancy appear to be altered as a result of natural selection [Labuda et al. 2007]. Natural selection is expected to affect the distribution of alleles around the selected site and thus of the resulting haplotypes captured in HACs. Looking for signatures of natural selection, we ask if the observed distribution of HACs significantly differs from the one expected under neutrality. Here we investigate the properties of HACs obtained under neutrality, under different demographic scenarios, with and without recombination, compared to the results obtained under natural selection. To detect differences in HACs distribution, we use principally the Svd-statistic. Svd can be calculated at each site by splitting in two sub-sets depending on the present or absent of the reference allele. Then, we subtract the variance of HAC distribution of these two sub-sets and we finish by multiplying the derived frequency at this site. Population samples were generated under the coalescence model using the ms program of Hudson [2002], assuming neutrality, recombination and different demographic scenarios including population bottlenecks and expansions. In turn, the SelSim program of Spencer and Coop [2004] was used to create samples with mutations under natural selection. We examined the distribution of HACs in the context of different demographic histories and have seen that HACs react in opposite way depending if there is a bottleneck or an exponential growth. Furthermore, when we test different recombination rates, we have seen that a high constant recombination rate diminishes variance of HAC distribution. As another aims, we tested the variation of HACs, when subjected to natural selection, allow to obtain the specific signal of the selection. All these simulations are generated for both constant number of polymorphism and constant population mutation rate. We also compared the performance of Svd with classical summary statistics such as Tajima's D and Fay & Wu's H. Svd seem to be more sensible to intermediate frequency selected alleles. HAC and Svd-statistic appear as particularly robust characteristics that can be used to visually examine and statistically assess genetic loci that evolve under selection constraints as compared with neutral expectation. (Research supported by FQRNT and a CIHR Strategic Training Program in Bioinformatics)

Affiches Ph.D.

Ph.D. 1 : Malika Aid :: Université de Montréal

DNA motif discovery approach adapted to ChIP-chip data

Transcription factors play an important role in various biological processes such as differentiation, cell cycle progression and tumorigenesis. They regulate gene transcription by binding to specific DNA sequences (cis-regulatory elements). Identifying these cis-regulatory elements is a crucial step in the understanding of gene regulatory networks. The recent developments in genomic technologies such as DNA microarrays and Chromatin immuno-precipitation followed by microarray hybridization (ChIP-chip) have made possible the whole genome characterization of TFs binding sites (TFBS) and allowed the development of several computational DNA motif discovery tools. Although these various tools are widely used and have led to the discovery of novel motifs, in practice none of them have proven to be efficient in control data sets due to a high rate of false positive and false negative predictions.

DNA motif discovery tools use different strategies to extract and represent the motif patterns: Enumerative or alignment-based approaches. Each of them uses a specific scoring function to evaluate motifs and report those having the highest scores compared to a reference data set. The main drawback of these scoring functions is the fact that some motifs occurring ubiquitously in the genome are scored very highly, despite not being real enriched (false positive predictions). Consequently, real enriched motifs with low scores are penalized (false negative predictions).

Analyses conducted on simulated and ChIP-chip data using different tools: Mdmodule (enumerative algorithm, MEME (Expectation and maximization algorithm), and MotifSampler (Gibbs algorithm), have shown that DNA motif discovery tool scoring functions do not represent the observed characteristics of TFBS in ChIP regions. For example they do not take into account that motifs representing real binding sites are more likely to reside near the center of the ChIP regions. Our results showed that these scoring functions are not adapted to the discovery of TFBS in ChIP-chip data.

We propose to implement two new scoring functions: motif positional bias and motif group specificity that take into consideration the characteristics of the distribution of TFBS in the ChIP regions. Motif positional bias measures how a given motif is distributed across the ChIP regions. It is expected that true binding sites will be enriched in specific positions (peak in the ChIP central regions) compared to what is expected by chance (background data set). Group specificity score, is a measure of how a given motif targets the set of Chip regions. We expect that true TFBS will be distributed evenly throughout the ChIP sequences and are clearly more frequent compared to a reference data set.

We applied these scoring functions on a simulated data set and on real ChIP data set. The results show that our approach enhances the DNA motif discovery tools predictions and significantly reduce the rate of false positive predictions.

Ph.D. 2 : Mathieu Courcelles :: Université de Montréal

Digging into the phosphoproteome – confident phosphorylation sites localization and kinetic profiling

Remarkable advances in high resolution mass spectrometry at high throughput rate and affinity media have facilitated large-scale phosphoproteome analyses in support to molecular and cellular biology projects. From a single experiment, hundreds to thousands phosphorylation sites can be identified and precisely located on the protein structure. These developments had significant impact on the study of global cell signaling events in response to chemical stimulation and on the design of specific kinase inhibitors for cancer drug therapies. In view of the wealth of information generated by these large scale phosphorproteomics experiments, novel and improved bioinformatics tools are now required to profile changes in phosphorylation and associate interacting partners involved in specific signaling cascade events. To this end, we developed a bioinformatics platform called ProteoConnections tailored to address the pressing needs of phosphoproteomics analyses. The deployed data processing strategy focused on generating dataset with low false positives identification rate, an important consideration because these data are the starting point for downstream experiments like site directed mutagenesis which is labor intensive task. First, we used the target/decoy search method to obtain a false positive estimation of identified peptides. Second, because poor fragmentation of phosphopeptides which prevent in some cases the precise localization of the site, a probabilistic confidence value is reported. The platform was conceived to store into a relational database all phosphorylation site identifications to collect all evidences from LC-MS/MS and allow fast interaction with other analysis tools. Known and characterized phosphorylation sites (in vivo function, associated kinase) from Phospho.ELM, the most comprehensive phosphorylation database, and Swiss-Prot are incorporated to conveniently track previous records. Many existing tools have been interfaced with the platform for specific phosphoproteome analysis. PPSP and NetworKIN, phosphorylation sites predictors, are used to identify potential kinases for identified sites. Overrepresented phosphorylation motifs in the dataset are searched with Motif-X. Motifs from uncharacterized kinases can be discovered from this approach. Finally, to gain further knowledge and have a global overview of phosphoproteome, the dataset of phosphoproteins are mapped on protein-protein interaction network from STRING database using Cytoscape. The application of these tools will be demonstrated with preliminary data on phosphoproteome kinetic profiling of IEC6 rat cells upon ERK pathway inhibition.

Ph.D. 3 : Claudia Kleinman :: Université de Montréal

Protein structural representations for evolutionary analysis

The influence of tridimensional protein structure on sequence evolution has only been studied implicitly, since most evolutionary models used in phylogenetics today make the assumption of independence between sites. Models that explicitly account for site-interdependencies resulting from protein terciary structure have been recently developed. These models are based on statistical potentials, pseudo-energy scores that measure the compatibility of a sequence for a given structure. In a recent work, we proposed a new framework to devise statistical potentials suited for this application in evolutionary biology, using a simplified representation of the structure that included amino acid contacts and solvent accessibility information. Here, we explore different forms of statistical potentials, based on different structure representations, so as to study which structural elements best explain features of protein evolution.

Ph.D. 4 : Louis-Philippe Lemieux-Perreault :: Université de Montréal

Variabilité dans la détermination du nombre de copies des polymorphismes de nombre selon le degré d'apparentement des sujets

Les polymorphismes génétiques d'un seul nucléotide (de l'anglais Single Nucleotide Polymorphism ou SNP) sont maintenant grandement utilisés dans le cadre d'études génétiques à visées étiologiques pour traits génétiquement complexes. Ils ont l'avantage d'être fréquents dans le génome, assurant une bonne couverture de ce dernier. Par contre, ils sont peu informatifs compte tenu de leur nature dichotomique, c'est-à-dire qu'ils ne possèdent que deux allèles. Depuis quelque temps, beaucoup d'attention est portée sur les polymorphismes de nombre (de l'anglais Copy Number Polymorphism ou CNP). Ceux-ci, sont moins fréquents dans le génome mais ils possèdent plusieurs variations possibles et apportent un degré d'information complémentaire à celui des SNP lors d'études génétiques. Les différentes plateformes d'analyses génétiques (Affymetrix et Illumina) offrent des micropuces permettant le génotypage de plus de 900 000 SNP et plus de 940 000 sondes pour les CNP]. Beaucoup de logiciels existent maintenant afin de procéder au groupement des génotypes (SNP calling). Tel n'est pas le cas pour la détermination du nombre de copies (basée sur l'intensité logarithmique). Dernièrement, l'institut Broad a développé une suite de logiciels, nommée BirdSuite, possédant entre autre un logiciel permettant de catégoriser les CNP. En considérant le fait que le groupement des SNP peut varier en fonction des caractéristiques de la population à l'étude (sujets apparentés ou non, taille de l'échantillon, etc.), nous posons l'hypothèse que la détermination du nombre de copie peut elle aussi, varier en fonction de ces mêmes caractéristiques. Cette variation peut avoir un impact important sur les résultats des études d'association et de liaison génétique et il est important d'évaluer et de caractériser cette variation. Pour la présente étude, nous avons utilisé des données génétiques provenant d'individus génotypé à l'aide de puces Genome-Wide Human SNP Array 6.0 d'Affymetrix dans le cadre d'une étude sur l'obstruction du débit ventriculaire gauche (Left Ventricular Outflow Tract Obstruction ou LVOTO). Cette étude a pour but de comparer les résultats obtenus suite à la détermination du nombre de copies à l'aide du logiciel Canary à partir des données brutes sur deux groupe de sujets, soit (1) sur une sélection d'individus issus de familles multi-générationnelles et (2) sur des individus sans liens de parentés. Les résultats seront présentés et discutés.

Ph.D. 5: Véronique Lisi :: Université de Montréal

miRNAs and transcription factors in auto-regulatory loops

Micro-RNAs (miRNAs) down-regulate gene expression by binding to the 3'-untranslated region of their target. They act as fine-tuner of gene expression and are themselves regulated similarly to any other Pol-II dependent gene. Some miRNAs are involved in auto-regulatory loops with transcription factors (TFs). Here, we show that using publicly available informations, we can predict auto-regulatory loops between miRNAs and TFs. We experimentally validated two such loops in support of our predictions.

Ph.D. 6 : Karine St-Onge :: Université de Montréal

A Quantitative Structure-Activity Approach Applied to the Determination of Key RNA Functional Structural Features Without Explicit Representation of Involved Co-Factors

Recently, modeling RNA 3-D structure from sequence data has been made easier than ever thanks to the MC-Fold and MC-Sym pipeline. We can now produce quickly accurate 3-D structures of many sequences. This is a critical step in the QSAR (Quantitative Structure-Activity Relationships) approach. From knowing the activity rates of a set of functionally-related sequences and their 3-D structures, a comparative approach allows us to identify physicochemical features that are important for function without explicit representation of all involved co-factors. Here, we present a computational method to identify functional structural features from a set of active and inactive sequences using QSAR. Our method offers three analysis modes: exposed atoms, exposed donor/acceptor groups, and exposed atom charges. We exemplify the method and its modes to the identification of the important functional features of the Sarcin/Ricin motif from a set of viable and lethal mutations. We show how our results corroborate with the data obtained by Correll et al.

Affiches post-doctorat

PD 1 : Ali Mokdad :: Université de Montréal

Three-dimensional modeling of human precursor MicroRNAs

There are about 700 human precursor microRNA (pre-miRNA) entries in miRBase database (1), none of which with a known three-dimensional (3D) structure. It is not even conceivable to determine experimentally the structure of each of them. Our aim here is to determine a viable and accurate 3D model for each human pre-miRNA, and then to extract the 3D rules that allow these miRNAs to function in vivo. It is important to remember that at first glance pre-miRNAs seem like regular RNA helices with uninteresting 3D structural features, but in fact they are seeded with specifically positioned and oriented non-canonical interactions and bulges that can only be studied and characterized in 3D. Recently, our lab released a set of software tools that efficiently and accurately predict RNA 3D models from primary sequence data. With these tools we obtain a set of solutions that represent the 2D and 3D possibilities an RNA sequence can fold into. The problem is then reduced to identifying the best models from the pool of results. To achieve this we implement known structural constraints derived from the only crystal structure of Dicer with a modeled double stranded RNA helix docked with it. Based on this Dicer structure, the range of lengths of mature miRNA sequences, and the statistics of preliminary 2D and 3D structures of pre-miRNAs that we have already determined, we hypothesize that "human pre-miRNAs do not fold in an identical fashion to form a rigid stem with exact length, but rather fold according to several possible helical templates that range in size between 18 and 22 base or base pair steps between Drosha and Dicer scissile phosphates". Dicer is thought to be flexible enough to accommodate and cleave the range of these sizes. To identify the natural helical template for each pre-miRNA, we are currently folding each of them in 3D according to the distance constraints of each of five predetermined templates. We will then select the one template that produces most low free-energy structures. Pre-miRNAs that belong to each template will be structurally aligned, allowing for the determination of their endogenous structural design rules. An application of this project is to use the obtained 3D rational design rules to create therapeutic miRNAs with high efficiency and low toxicity compared to current artificial miRNAs.

Merci à notre commanditaire !