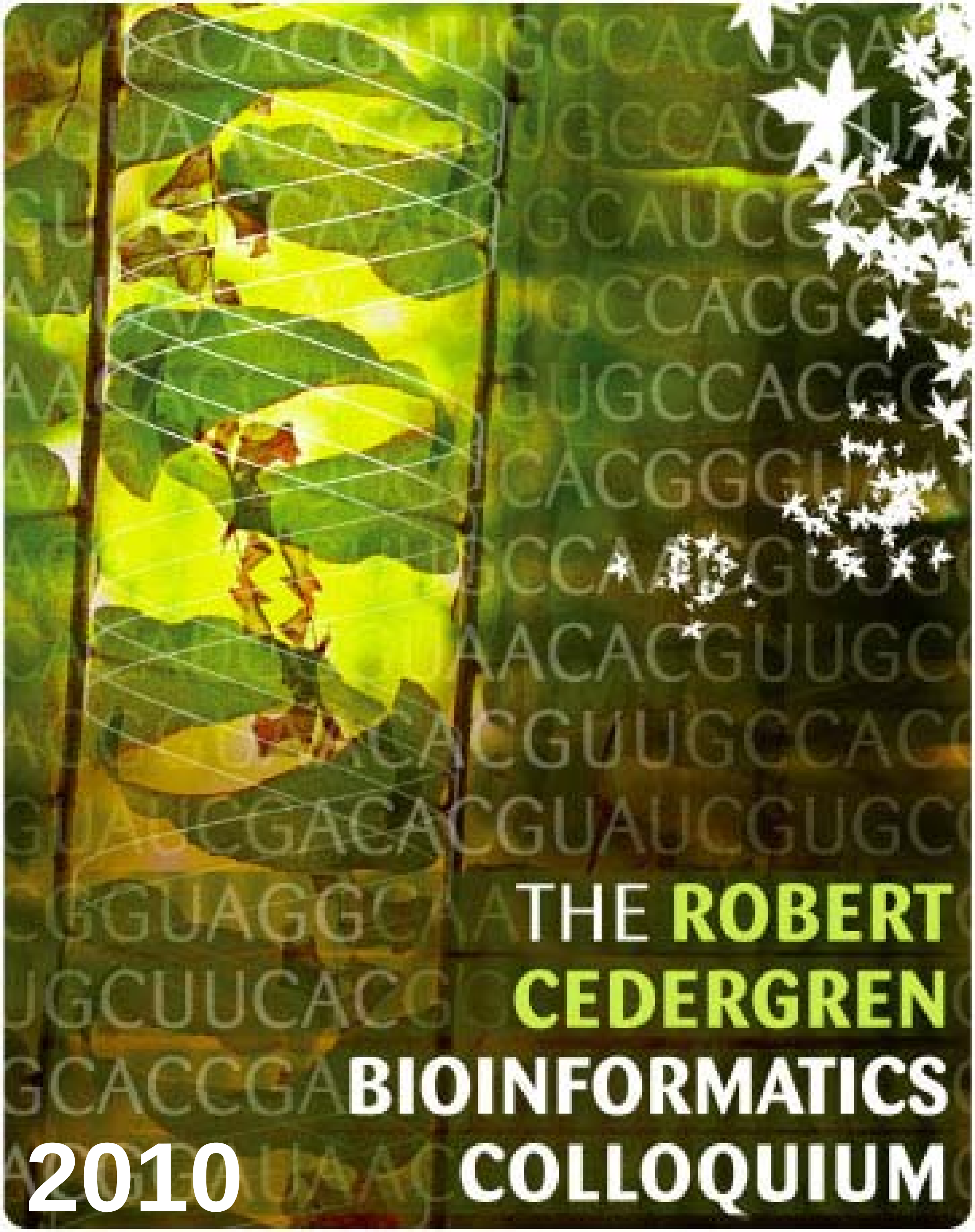

THE ROBERT CEDERGREN BIOINFORMATICS COLLOQUIUM

2010

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

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

GE Healthcare
Canada

sité
Montréal

Welcome to the 7th annual Robert Cedergren Colloquium in Bioinformatics !

This Colloquium is an annual event gathering the universitary community working in Bioinformatics. The main purpose is to share the latest advancement in Bioinformatics by posters and talks, and to demonstrate the increasing role of Bioinformatics for Life sciences in general and Health research in particular.

This year's invited keynote speakers are

Michael Hallett, McGill Centre for Bioinformatics, School of Computer Science, McGill University, Montreal, QC, Canada

Peter Karp, Artificial Intelligence Center, Director of Bioinformatics Research Group, SRI International, Menlo Park, CA, USA

Normand Mousseau, Département de Physique, Centre Robert-Cedergren in Bioinformatics and Genomics, GÉPROM, Université de Montréal, Montreal, QC, Canada

Alfonso Valencia Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.

Graduate students are eligible for best-presentation awards in the following categories:

Category	Best Talk	Best posters
M.Sc	$ 1000	$ 500
Ph.D	$ 1000	$ 500

Enjoy the Colloquium !

Gertraud Burger, Ph.D.
Leader, bioinformatics training programs
Université de Montréal

Bienvenue au 7e 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'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 septième edition du colloque, les conférenciers invites sont :

Michael Hallett, McGill Centre for Bioinformatics, School of Computer Science, McGill University, Montreal, QC, Canada

Peter Karp, Artificial Intelligence Center, Director of Bioinformatics Research Group, SRI International, Menlo Park, CA, USA

Normand Mousseau, Département de Physique, Centre Robert-Cedergren in Bioinformatics and Genomics, GÉPROM, Université de Montréal, Montreal, QC, Canada

Alfonso Valencia Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.

Des prix seront décernés dans les categories suivantes :

Catégorie	Meilleures présentations orales	Meilleures affiches
M.Sc	$ 1000	$ 500
Ph.D	$ 1000	$ 500

Un excellent colloque à tous et à toutes !

Gertraud Burger, Ph.D.
Coresponsable des programmes de bio-informatique
Université de Montréal

Table des matières / Contents

Comités / Committees

Comité scientifique / Scientific Committee

Gertraud Burger, Biochemistry, UdeM
Mathieu Blanchette, Comp. Sci., McGill
B. Franz Lang, Biochemistry, UdeM
Hervé Philippe, Biochemistry, UdeM
Jacques Simard, CHUQ, Université Laval
Daniel Sinnett, Ste-Justine Hosp., UdeM
Elisabeth Tillier, UofToronto
Marcel Turcotte, Info.Technology, U.Ottawa

Modérateurs / Session Chairs

B. Franz Lang, Biochemistry, UdeM
Nicolas Lartillot, Biochemistry, UdeM

Jurys

MSc-posters	(1)
MSc-talks	(2)
PhD-posters	(3)
PhD-talks	(4)

Juges / Referees (Presidents in bold)

Henner Brinkmann, Biochemistry, UdeM	(1)
Serguei Chteinberg, Biochemistry, UdeM	**(1)**
Arnaud Droit, Université Laval	(2)
Nadia El-Mabrouk, DIRO, UdeM	(1)
Sylvain Foisy, Montreal Heart Institute	(2)
Claudia Kleinman, McGill University	**(2)**
B. Franz Lang, Biochemistry, UdeM	(3)
Nicolas Lartillot, Biochemistry, UdeM	(4)
Vladimir Makarenkov, Comp.Sci., UQAM	**(3)**
Alejandro Murua, Math & Stat., UdeM	(3)
Emmet O'Brien, Comp. Sci., Concordia	(4)
Hervé Philippe, Biochemistry, UdeM	**(4)**

Comité organisateur / Organizing Committee
(Robert-Cedergren Centre)

Gertraud Burger
B. Franz Lang
Sandrine Moreira

Bénévoles / Volunteers
(Dept. Biochemistry, Robert-Cedergren Centre)

Natacha Beck
Georgette Kiethega
Philipe Lampron
Elaine Meunier
Ioana Minoiu
Audrey Noel
Yifei Yan

Commanditaires / Sponsors

Nous remercions le généreux support de / We thank for generous support from the

Canadian Institute of Health Research (CIHR)
Vice-Rectorate of Research, Université de Montréal
Faculté des arts et des sciences, Université de Montréal
Département de biochimie, Université de Montréal
GE Healthcare Canada
Cedergren family, including

- Henrietta Jonas-Cedergren, Montreal, Canada
- Patricia Cedergren, Laguna Niguel, California, USA
- Janet Cedergren Gilbertson, Coon Rapids, Minnesota, USA
- Barbara Cedergren Piculell, Burnsville, Minnesota, USA
- Margaret Cedergren Vaughan, Chebeague Is., Maine, USA.

Renseignements généraux / General information

Accueil / Registration

L'accueil des participants se fera au Hall d'honneur, Pavillon Roger-Gaudry (2900 Edouard-Montpetit, sous la tour, en haut de l'escalier), le 1er et 2 novembre dès 8h15.

The registration desk is located in the Hall d'honneur, Pavillon Roger Gaudry (2900 Edouard-Montpetit; under the tower, accessible via the monumental stone stairs), and open on 1-2 November, from 8:15 am.

Présentations orales / Talks

Les presentations orales auront lieu dans la sale M-415, à côté du Hall d'honneur.

Oral presentations will take place in Room M-415, next to the Hall.

Affiches / Poster sessions

Les affiches seront exposées dans le Hall d'honneur. Les auteurs devront être devant leur affiche pendant les sessions d'affiches, prière de mettre un mot d'absence indiquant le moment votre retour si vous devez vous absenter.

Posters will be displayed in the Hall. Poster authors should be at their posters during all poster sessions and when absent, post a note indicating when they are back.

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

Les pauses santé, les lunchs et le cocktail seront servis dans le Hall d'honneur.

Coffee breaks, lunch and cocktail will be served in the Hall d'honneur.

Tableau des messages / Message board

Pour afficher un message, utilisez le tableau près du bureau d'accueil.

For posting messages, use the board close to the registration desk.

Accès internet / Internet access

Nous essairons de fournir l'accès de participants de conférence au réseau WiFi pendant le colloque.

We are attempting to provide conference participants access to the WiFi network during the colloquium.

Programme / Program

Monday November 1		
Session 1 Chair: BF Lang	Time	All Talks in room M-415, Pavillon Roger Gaudry, Université de Montréal
	08:30-09:15	**Registration:** Hall d’honneur, Pavillon Roger Gaudry
	09:15-09:30	**Colloquium opening**: Gertraud Burger, RC-Centre, Université de Montréal
	09:30-10:30	**Keynote**: Peter Karp, Artificial Intelligence Center, SRI, Menlo Park, CA, USA *“Exploration and refinement of the HumanCyc human metabolic network model”*
	10:30-10:45	**Coffee break**
	10:45-11:30	**Poster Session A** (Hall d’honneur)
	11:30-12:00	**Talk** MSc: Carl Song, Hospital for Sick Children, University of Toronto *“Characterizing apicomplexan parasite metabolism by flux balance analysis of* Toxoplasma gondii”
	12:00-12:30	**Talk** MSc: Anna van Weringh, Department of Biology, University of Ottawa *“Removing falsely predicted tRNA genes from eukaryotic species by RNA phylogenetics”*
	12:30-13:00	**Talk** MSc: Raphael Poujol, RC-Centre, Université de Montréal *“A Bayesian phylogenomic approach for detecting longevity-related genes in mammals”*
	13:00-13:45	**Lunch** (Hall d’honneur)
Session 2 Chair: Nicolas Lartillot	13:45-15:30	**Poster Session B** (Hall d’honneur) - **[15:15 Jury 1 convenes]**
	13:45-15:00	**Coffee Break** (Hall d’honneur)
	15:30-16:00	**Talk** MSc: Jean-Christophe Grenier, RC-Centre, Université de Montréal *“Accuracy of archaeal phylogeny: do horizontal gene transfers or models of sequence evolution matter most?”*
	16:00-16:30	**Talk** MSc: Yves Gagnon, DIRO, Université de Montréal *« Reconstruction of ancestral genome subject to whole genome duplication, speciation, rearrangement and loss”*
	16:30-17:00	**Talk** PhD: Julie Hussin, Hôpital Ste-Justine, Université de Montréal *Age and sex-specific effects acting on recombination rates in humans*
	17:00-17:15	**Break - [Jury 2 convenes]**
	17:15-18:15	**Keynote**: Alfonso Valencia, CNIO National Cancer Research Centre Madrid, Spain *“The secret life of proteins”*

TUESDAY NOVEMBER 2		
Time		All talks in room M-415 Pavillon Roger Gaudry, Université de Montréal
Session 3 Chair: N Lartillot	08:30-09:15	**Registration** (Hall d'honneur)
	09:15-10:15	**Keynote**: Normand Mousseau, Dept. Physique & RC-Centre, Université de Montréal *"Simulation of protein flexibility and aggregation dynamics"*
	10:15-10:45	**Coffee break** (Hall d'honneur)
	10:15-11:30	**Poster Session C** (Hall d'honneur)
	11:30-12:00	**Talk** PhD: Mathieu Rousseau, Centre for Bioinformatics, McGill University *"Markov Chain Monte Carlo computational analysis of chromosome conformation capture Carbon copy data"*
	12:00-12:30	**Talk** PhD: Olivier Tremblay-Savard, DIRO, Université de Montréal *"Advances on genome duplication distances"*
	12:30-13:00	**Talk** PhD: Béatrice Roure, RC-Centre, Université de Montréal *"Are incomplete data matrices affecting negatively the accuracy of phylogenomics?"*
	13:00-13:45	**Lunch** (Hall d'honneur)
Session 4 Chair: BF Lang	13:45-16:00	**Poster Session D** (Hall d'honneur)
	13:45-15:30	**Coffee break** (Hall d'honneur)
	16:00-16:30	**Talk** PhD: Mathieu Lavallée-Adam, Centre for Bioinformatics, McGill University *"Modeling contaminants in TAP-MS/MS experiments"*
	16:30-17:00	**Talk** PDF: Magali Michaut, Centre for Cell. & Biomolec. Res, University of Toronto *"Bringing order to disorder: genomic analysis uncovers three distinct forms of protein disorder"*
	17:00-17:15	**Break - [Jurys 3,4 convene]**
	17:15-18:15	**Keynote**: Michael Hallett, Centre for Bioinformatics, McGill University *"A systems approach to understanding breast cancer"*
	18:15-18:45	**Awards** (Henrietta Jonas-Cedergren, Gertraud Burger)
		Colloquium closure (Gertraud Burger)
	18:45-21:00	**Cocktail** (Hall d'honneur)

Résumés / Abstracts

Les résumés sont regroupés en trios sections, 1. Conférenciers (K), 2. Présentations orales (T), et 3. Affiches (P). Dans chaque section, les résumés sont en ordre alphabétique par le nom de famille qui est souligné.

Les étudiants gradués sont éligibles pour les prix dans les catégories Affiches-MSc (P-MSc), Présentations orales-MSc (T-MSc), Affiches-PhD (P-PhD) et Présentations orales-PhD (T-PhD).

The abstracts are grouped into three sections, 1. Keynotes (K), 2. Talks (T), and 3. Posters (P). Within these sections, abstracts are ordered by the presenter's family name, which is underlined.

Graduate students are eligible for awards in the categories Posters-MSc, Talks-MSc, Posters-PhD and Talks-PhD, and the corresponding abstracts are labelled accordingly (P-MSc, T-MSc, T-PhD, or P-PhD).

Award candidates MSc level		
Name	*Poster*	*Talk*
Bastien D	#2	
Boufaden A	#6	
Fabre L	#7	
Gagnon Y		X
Grenier JC		X
Lalonde E *CANCELLED*	#10	
O'Reilly P	#18	
Poujol R		X
Ragonnet-Cronin M	#20	
Song C		X
Theroux JF	#25	
van Weringh A		X
Vello E	#26	

Award candidates PhD level		
Name	*Poster*	*Talk*
Aid M	#1	
Boc A	#4	
Bokov K	#5	
Fourati S	#8	
Hussin J		X
Ishi T	#9	
Lavallée-Adam M	#11	X
Lord E	#13	
Malekpour SA	#14	
Moreira S	#16	
Nikbakht H	#17	
Parto S	#19	
Roure B		X
Rousseau M	#21	X
Shateri Najafabadi H	#22	
St-Onge K	#23	
St-Pierre JF	#24	
Tremblay-Savard O		X

Conférences / Keynote Abstracts

(K) **A systems approach to understanding breast cancer**

Michael **Hallett**

Centre for Bioinformatics, McGill University, Montreal, QC, Canada

It is increasingly evident that breast cancer outcome is strongly influenced by signals emanating from tumor-associated stroma. However, little is known about how gene expression changes in this tissue affect tumor progression. In this talk, we compare gene expression profiles from laser capture-microdissected tumor-associated versus matched normal stroma, and derive transcriptional profiles strongly associated with clinical outcome. We present a stroma-derived predictor that generates new information to stratify disease endpoint, independent of standard clinical prognostic factors and previously published predictors. Our predictor selects poor-outcome patients from multiple clinical subtypes, including node-negative patients, and predicts outcome in multiple published expression datasets generated from whole tumor tissue. Our predictor has increased accuracy compared to previously published predictors, and prognostic accuracy increases when these predictors are integrated using graphical models. Genes represented in the stroma-derived predictor reveal the strong prognostic capacity of differential immune responses as well as angiogenic and hypoxic responses.

The computational and statistical aspects underpinning this work are built upon a new approach to analyzing gene expression data that in some sense is "orthologonal" to traditional clustering based tools, and is general in the sense that a wide range of data types can be easily integrated into the system.

(K) **Modeling the Human Metabolic Network in HumanCyc**

Peter D. **Karp**

Artificial Intelligence Center, SRI International, Menlo Park, CA, USA

The HumanCyc database describes 247 human metabolic pathways and their associated enzymes, reactions, metabolites, and transporters. HumanCyc has been created through a combination of computational inferences and manual curation. This talk will describe the methodology by which HumanCyc is developed, and survey its knowledge content. We will discuss computational tools used to verify the accuracy and completeness of the database, including identification of dead-end metabolites and reachability analysis. The talk will also describe application of HumanCyc to human metabolomics studies underway by the Metabolomics Network.

(K) **Simulation of protein flexibility and aggregation dynamics**

Normand **Mousseau**

Département de Physique, Centre Robert-Cedergren in Bioinformatics and Genomics, GÉPROM, Université de Montréal, Montreal, QC, Canada

Insoluble amyloid fibrils are associated with a number of degenerative diseases such as Alzheimer's, Parkinson's and Huntington's diseases. While the exact structure of these assemblies is still unknown, it has become clear, in the last decade, that small soluble intermediate structures are much more toxic than the fully fibrils. Because of the inherent kinetic dynamical nature of these small aggregates, however, it has been very difficult to obtain clear experimental results, and computer simulations have emerged as an essential tool in the study of the onset of amyloid formation. In this talk, I will discuss the state of the art in simulating these systems, with an emphasis on the results obtained by our group.

This work was done in collaboration with Sébastien Côté, Jessica Nasica-Labouze, Rozita Laghaei from the Université de Montréal as well as Giorgio Colombo (CNRC, Milan), Philippe Derreumaux (IBPC, Paris) and Guanghong Wei (Fudan, Shanghai).

(K) **The secret life of proteins**

Alfonso **Valencia**

Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO) Madrid, Spain

My intention with this talk is to cover on-going developments in the field of protein binding sites and protein interactions and, perhaps more importantly, to point to the many interesting and still widely open problems in this fascinating area of research.

In the first part of the talk I will describe our recent efforts to predict and understand protein binding sites and protein interactions. In the second part I will introduce the new challenges that splicing creates for the interpretation of functional information. Finally, I will discuss the use of protein interaction networks as an additional possibility for the exploration of the functional space and its applications for the interpretation of cancer genome data.

My group is particularly interested the development of methods for the prediction of protein functional sites, i.e. interactions with ligands and other proteins, based on ideas derived from the field of evolution to detect key regions for the organization of protein binding sites (Rausell et al., PNAS 2010) and for the prediction protein interactions at the genomic scale (Juan et al., PNAS 2008). Regarding the validation of predicted interaction and interaction sites we have developed specific resources (Lopez et al., NAR 2007) and introduced standards for the evaluation of the predictions (Lopez et al., Proteins 2009, Ezkurdia et al., Brief Bioinf 2009). Furthermore, we are using information directly extracted from the primary publications to validate the predictions. Along this line we have recently organized the BIOCREATVE II.5 challenge (http://www.biocreative.org) in collaboration with the MINT databases. 15 teams participated in this global evaluation of the capacity of information extraction servers to reproduce the information on protein interactions provided by authors and database curators. The promising results clearly pointed to capacity of the automatic systems to contribute to the generation of Structure Digital Abstracts in collaboration with authors, databases and publishers (Leitner et al., Nat Biotech 2010).

In the second part of the talk, I will put this general methodology for the study of functional sites in the context of the analysis of the possible function of the potentially produced splicing isoforms following our initial work in this area (Tress et al, PNAS 2007, Bioinformatics 2008) and the on-going efforts to assess the actual presence of protein splice isoforms in cellular systems (Tress et al., Genome Biology 2008and 2010 submitted).) This collection of evidences suggests that the actual panorama of protein functions in the cell, and the contribution of splice variants, can be more complex than anticipated.

Finally, I will introduce our recent work addressing the use of protein interaction networks in the analysis of functional information (Baudot et al., Genome Biology 2009), starting with some new exciting results that challenge some of the prevailing views on protein interactions (Wass et al., 2010 submitted), and ending with our current approach to the interpretation of cancer genome data (potential cancer related genes) based on the use of interaction networks and biomedical meta-information (Baudot et al., EMBO Rep 2010).

Présentations orales / Talk Abstracts

(T-MSc) **Reconstruction of ancestral genome subject to whole genome duplication, speciation, rearrangement and loss**

Denis Bertrand[1], Yves **Gagnon**[1], Mathieu Blanchette[2], and Nadia El-Mabrouk[1]

[1] Département d'Informatique et Recherche Opérationnelle, Robert Cedergren Centre in Bioinformatics and Genomics, Université de Montréal, Montreal, Canada
[2] McGill Center for Bioinformatics, McGill University, Montreal, QC, Canada

Whole genome duplication (WGD) is a rare evolutionary event that has played a dramatic role in the diversification of most eukaryotic lineages. Given a set of species known to have evolved from a common ancestor through one or many rounds of WGD together with a set of genome rearrangements, and a phylogenetic tree for these species, the goal is to infer the pre-duplicated ancestral genomes.

We use a two step approach: (i) compute a score for each possible ancestral adjacency at each internal node of the phylogeny; ii) combine adjacencies to form ancestral chromosomes. The main contribution of our method is the computation of a rigorous score for each potential ancestral adjacency (a,b) reflecting the maximum number of times a and b can be adjacent in the whole phylogeny, for any setting of ancestral genomes.

We first apply our method on simulated datasets and show a high accuracy for adjacency prediction for scenarios both with and without WGD events. In scenarios whithout WGD events, our algorithm outperforms a similar method by Ma et al. [1]. We then infer the pre-duplicated ancestor of a set of 11 yeast species and compare it to a manually assembled ancestral genome obtained by Gordon et al. [2].

References

1. J. Ma et al. Reconstructing contiguous regions of an ancestral genome. Genome Research, 16:1557- 1565, 2007
2. J.L. Gordon, K.P. Byrne, and K.H. Wolfe. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern saccharomyces cerevisiae genome. PloS Genetics, 5(5), 2009.

(T-MSc) **Accuracy of archaeal phylogeny: do horizontal gene transfers or models of sequence evolution matter most?**

Jean-Christophe **Grenier**[1] , Marie-Ka Tilak[2], Nicolas Lartillot[1], Henner Brinkmann[1] and Hervé Philippe[1]

[1] Robert Cedergren Centre in Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, Montréal. QC, Canada.
[2] Laboratoire de Paléontologie, Phylogénie & Paléobiologie, Institut des Sciences de l'Evolution (UMR 5554 CNRS), Université Montpellier II, Montpellier, France.

Horizontal gene transfer had been demonstrated to play an important role in the evolution of prokaryotes. Their negative impact on phylogeny was the subject of a heated debate, with some authors even proposing that the concept of a species tree for prokaryotes should be abandoned. Since stable and functional horizontal transmissions appear to be by far rarer than vertical transmissions (tens versus billions), the phylogeny of prokaryotes does contain a major part of the historical signal. However, the cumulative effect of horizontal gene transfers is non-negligible and can potentially affect phylogenetic inference in a negative way. Therefore, most researchers base their phylogenetic inference on a low number of rarely transferred genes such as ribosomal proteins, but they assume the selection of inference methods as less important, this despite the fact that it has been shown that the model of sequence evolution is of prime importance for much less deep divergences, e.g. like animals.

Here, we used a combination of simulations and of real data from Archaea to study the relative impact of horizontal gene transfers and of the inference methods on the phylogenetic accuracy. Our simulations prove that (1) horizontal gene transfers have a limited impact on phylogeny, assuming a realistic rate and (2) the supermatrix is much more accurate than the supertree approach. We were trying to improve several archaeal phylogenies coming from the recent literature, by applying more complex models of sequence evolution. We observed that more complex models of evolution not only have a better fit to the data, but can also have a direct impact on different phylogenetic groups and on the robustness of the tree. In particular, our results are in contradiction to recent publications proposing that the Thaumarchaeota are at the base of the archaeal tree.

(T-PhD) **Age and sex-specific effects acting on recombination rates in humans**

Julie **Hussin**[1], Marie-Hélène Roy-Gagnon[3], Roxanne Gendron[3], Gregor Andelfinger[2,3], and Philip Awadalla[1-3]

[1] Département de Biochimie, Faculté de Médecine, Université de Montréal, Montreal, QC, Canada
[2] Centre de recherche de l'hôpital Ste-Justine, Université de Montréal, Montreal, QC, Canada
[3] Département de Pédiatrie, Faculté de Médecine, Université de Montréal, Montreal, QC, Canada

In humans, chromosome-number abnormalities in offspring have been associated with altered recombination and increased maternal age. The underlying causes of the latter association remain unknown, therefore age-related effects on recombination are of major importance, especially in relation to the mechanisms involved in human trisomies. In rodent oocytes, the frequency of recombination events has been found to decrease with age. We focus here on determining whether recombination rate is related to the age of the mother in humans.

We localized crossovers at high resolution using a dense genome-wide SNP survey (6.0 Affymetrix platform) genotyped among French-Canadian multi-generation pedigrees, providing information about 195 maternal meiosis. Overall, we observed similar variation in fine-scale recombination rates and patterns as previously observed in Hutterites families (Coop et al. Science. 2008). However, contrary to what has been previously reported for humans (Kong et al. Nature Genetics. 2004), we observed that viable offspring of older mothers tend to have reduced recombination rates, in agreement with the findings in mouse and hamster.

The most pronounced effect is seen for mothers over the age of 30. The observation is a genome-wide effect but, among submetacentric chromosomes, the effect was significantly more pronounced, suggesting a subtelomeric effect. Furthermore, we observed that in females, recombination frequencies drop dramatically in subtelomeric regions. Finally, we propose a model that reconciles our findings with earlier reports that found associations between maternal age and recombination in trisomy cases.

(T-PhD) **Modeling contaminants in TAP-MS/MS experiments**

Mathieu **Lavallée-Adam**[1], Philippe Cloutier[2], Benoit Coulombe[2], and Mathieu Blanchette[1]

[1] McGill Centre for Bioinformatics and School of Computer Science, McGill University, Montréal, QC, Canada
[2] Gene Transcription and Proteomics Laboratory, Institut de recherches cliniques de Montréal, Montréal, QC, Canada

Identification of protein-protein interactions (PPI) by tandem affinity purification (TAP) coupled with tandem mass spectrometry (TAP-MS/MS) produces large datasets with high rates of false positives. This is in part because of contamination at the TAP level (due to gel contamination, non-specific binding to the TAP columns, insufficient purification, etc.).

We introduce a Bayesian approach to identify false-positive PPIs involving contaminants in TAP-MS/MS experiments. Specifically, we propose a confidence assessment algorithm that builds a model of contaminants using a small number of representative control experiments. It then uses this model to determine whether the Mascot score of a putative prey is significantly larger than what was observed in control experiments and assigns it a p-value and a false discovery rate.

We show that our method identifies contaminants better than previously used approaches and results in a set of PPIs with a larger overlap with databases of known PPIs. Our approach will allow improved accuracy in PPI identification while reducing the number of control experiments required.

(T) **Bringing order to disorder: genomic analysis uncovers three distinct forms of protein disorder**

Magali **Michaut** [2,3]*, J Bellay[1*], S Han[2,3]*, M Constanzo[2,3], BJ Andrews[2,-4], C Boone[1-4], GD Bader[2-5], CL Myers[1], PM Kim[2-5]

[1] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA
[2] Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada
[3] Banting and Best Department of Medical Research, University of Toronto, Toronto, ON, Canada
[4] Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
[5] Department of Computer Science, University of Toronto, Toronto, ON, Canada
* These authors contributed equally to this work

Motivation

Intrinsically disordered regions are widespread, especially in proteomes of higher eukaryotes. Intrinsically disordered proteins, which have a large fraction of disordered residues, have been associated with a large variety of functions and many diseases. However, a detailed understanding of the role of disordered regions has remained elusive. In this study, we help to shine some light on this by systematically distinguishing different types of intrinsic disorder using a novel analysis that leverages both comparative genomics and genetic interactions.

Methods

To investigate disordered proteins in the context of genetic interactions, we used the most comprehensive genetic interaction network for *S. cerevisiae* (Costanzo et al. 2010). To examine their evolutionary properties we investigated which disordered regions were also disordered in orthologous proteins across the yeast clade. We defined three distinct classes: regions of conserved disorder with quickly evolving sequences (flexible disorder), regions where disorder is conserved with highly constrained amino acid sequence (constrained disorder) and, lastly, non-conserved disorder.

Results

We observed that genes that have numerous genetic interactions often tend to encode proteins that have a higher percentage of disordered residues. We found that regions of conserved disorder are strongly predictive for harboring linear motifs and phosphorylation sites. Flexible disorder is closest to the canonical notion of protein disorder and is responsible for its association with signaling pathways and multi-functionality. Conversely, proteins high in constrained disorder are involved in RNA binding and protein folding. Non-conserved disorder appears to be largely non-functional sequence.

Discussion

We thus conclude that by analyzing evolutionary signatures of disordered regions we can distinguish three functionally distinct subdivisions that also correspond to biophysically distinct phenomena. (T-MSc) A Bayesian phylogenomic approach for detecting longevity-related genes in mammals.

(T-MSc) **A Bayesian phylogenomic approach for detecting longevity-related genes in mammals**

Raphael **Poujol**, Nicolas Lartillot

Robert Cedergren Centre in Bioinformatics and Genomics, Department of Biochemistry, Université de Montreal, Montreal, QC, Canada

Studies suggest that aging is due to oxidative stress, and that it is further aggravated by errors in proteins synthesis. Many genes have been proposed to play a role in the prevention of cell degeneration and premature aging. Assuming that those genes are subject to stronger selective pressure in long-lived species, we can rely on observed correlations between selection pressure and longevity across currently available mammalian genomes to screen for longevity-related candidate genes. The selective pressure on a gene can be estimated by ω, the ratio of on-synonymous (dN) to synonymous (dS) substitution rates over time (ω = dN / dS). Lower values of ω indicate a stronger selective pressure. Our key assumption is that genes for which ω and longevity have opposite variations (i.e. are negatively correlated) are likely involved in the regulation of aging.

As more generally in comparative biology, estimating correlations between related species should be made in a phylogenetic framework, in order to get rid of the confounding effects of evolutionary inertia. Accordingly, the main idea of my study is to model the correlated history of longevity and gene-specific selective pressure (ω) along the lineages of the mammalian tree using Brownian processes. The covariances and all the parameters of the model are estimated by Markov Chain Monte Carlo. The overall method is thus equivalent to a hierarchical Bayesian phylogenetic regression framework, and can be used to get new insights about the genes potentially involved in the evolution any continuous morphological or life-history trait.

(T-PhD) **Are incomplete data matrices affecting negatively the accuracy of phylogenomics?**

Béatrice **Roure** and Hervé Philippe

Robert Cedergren Centre in Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, Montreal, QC, Canada

Phylogenomics, the use of large-scale datasets, is becoming a standard approach to infer the Tree of Life, thanks to its large amount of phylogenetic signal. However, assembling a complete dataset with all species for all genes remains difficult: even with completely sequenced genomes, gene lost, xenologous copy or duplications may render the orthologous copy really absent. Moreover, data are often obtained through random sequencing of ESTs or targeted sequencing of PCR products, and since both approaches regularly fail to obtain the genes of interest, the frequency of missing data in phylogenomic dataset can become important (up to 95%). Although the literature on the impact of missing data is relatively abundant, almost all articles use simulations and focus on small datasets (typically 1000 positions). Most studies (especially from Wiens' group) indicate that missing data do not seriously decrease phylogenetic accuracy as long as a sufficient number of characters is available, but a recent study of Lemmon et al. (Syst Biol. 2009;58(1):130-45) suggests that phylogenetic inferences using probabilistic methods might be highly sensitive to missing data. Here, we used real datasets to investigate whether missing data affect negatively the accuracy of phylogenomics.

Starting from a complete alignment from Metazoa (39 species, 126 nuclear proteins, 29,715 sites), we introduced variable amounts of missing data from 20 to 80% in two clades (deuterostomes and protostomes), mimicking the distribution of gaps observed in EST-based datasets. As a control, the same proportion of positions was removed for all species, thereby creating smaller but complete alignments. Phylogenies were inferred by the maximum likelihood method with RAxML using a WAG+F+?4 model and by a Bayesian method with PhyloBayes using a CAT+?4 model. Phylogenomics appears to be highly robust to the presence of a large amount of missing data. Almost all nodes that are highly supported with the complete alignment remain highly supported even when 80% of the data are missing in one of the two major animal clades. The most notable exception is the position of the fast evolving Diptera, which progressively move to the base of hexapods, when missing data are introduced in protostomes, but neither when introduced in deuterostomes, nor when complete genes are removed. The long-branch attraction (LBA) artefact is therefore enhanced by the presence of missing data, a result we confirmed for the position of ctenophores with another dataset. However, the comparison of the two models of sequence evolution revealed that the choice of the model plays a much more important role in the reduction of the deleterious effect of LBA: the CAT+?4 model, which has a significantly better fit than the WAG+F+?4 model, more accurately locates the fast evolving taxa (nematodes, dipterans, platyhelminths, rotifers) and is less affected by the increasing amount of missing data.

In conclusion, we showed that missing data can have a negative impact on phylogenomic accuracy, when the problem of interest is difficult (short internal branch and/or accelerated evolutionary rate), but that the effect of the model of sequence evolution is much more important. As a result, even if additional studies on missing data are needed, future research should focus on improving tree reconstruction methods.

(T-PhD) **Markov Chain Monte Carlo computational analysis of chromosome conformation capture Carbon copy data**

Mathieu **Rousseau**[1], James Fraser[2], Josée Dostie[2], Mathieu Blanchette[1]

[1] McGill Centre for Bioinformatics, McGill University, Montreal, QC, Canada
[2] Department of Biochemistry and Rosalind and Morris Goodman Cancer Centre, McGill University, Montreal, QC, Canada

Objectives

Long-range interactions of enhancers and insulators with the transcription initiation machinery play an important role in regulating gene expression. These interactions are the consequence of the 3D conformation of chromatin within the cell. A technique called Chromosome Conformation Capture Carbon Copy (5C) is used to measure the interaction frequency (IF) between specific regions of the genome. Our goal is to use the IF data generated by 5C experimentation to computationally model and analyze three-dimensional chromatin structure.

Methods

A Markov chain Monte Carlo (MCMC) approach was used to generate a representative sample of structures from IF data using the Metropolis-Hastings algorithm. The structures were clustered using a hierarchical clustering method to identify structure subfamilies. Each structure subfamily was further analyzed using a minimum-weight clique finding heuristic to identify reliable substructures.

Results

A biological model of undifferentiated myeolomonocyte THP-1 cells treated with phorbol myristate acid (PMA) to induce differentiation to macrophage cells was used. 5C data was generated for the HoxA gene cluster before and after differentiation. Parallel MCMC runs on the same dataset produced structures that showed that our MCMC method mixes quickly and is able to sample from the posterior distribution of structures. Structures generated from IF data from before and after differentiation formed two separate groups with no inter-mixing when clustered.

Conclusion

The Markov chain Monte Carlo approach used to model three-dimensional chromatin structure from 5C interaction frequency data samples structures from the posterior distribution. Clustering of the structures generated from before and after myelonomonocyte differentiation reveals distinct structure subfamilies. These structure subfamilies correlate with changes in HoxA gene expression before and after differentiation. The transcription start sites for the genes with differential expression are contained within reliable substructures. These findings indicate that changes of the chromatin structure in genomic regions surrounding the transcription start site may play a role in regulating gene expression.

(T-MSc) **Characterizing apicomplexan parasite metabolism by flux balance analysis of *Toxoplasma gondii***

Carl **Song** and John Parkinson

Molecular Structure and Function, The Hospital for Sick Children, Toronto, ON, Canada

The increasing prevalence of infections involving apicomplexan parasites such as *Plasmodium*, *Toxoplasma*, and *Cryptosporidium* (causative agents of malaria, toxoplasmosis and cryptosporidiosis respectively) represents a significant global healthcare burden. Despite their significance, few treatments are available, and the situation further deteriorates with the emergence of new resistant strains of parasites. We postulate that parasites have evolved distinct metabolic strategies critical for growth and survival during human infections. We further hypothesize that the enzymes which undertake these critical functions represent potent virulence factors that form components of highly integrated metabolic networks. Unfortunately, current knowledge of the metabolic potential of apicomplexan parasites throughout the course of an infection is rudimentary at best.

In order to fully understand the complex parasite-host relationships and identifying those enzymes that mediate critical roles from a global "systems" perspective, a fully characterized metabolic network of the experimentally amenable model apicomplexan - *Toxoplasma gondii* - has been reconstructed through extensive curation of available genomic and biochemical data. Using a sophisticated mathematical modeling framework, we are currently applying flux balance analysis to explore the metabolic potential of the parasite, and to identify highly enzymes that mediate critical roles for its growth. Preliminary results show that *T. gondii* incorporates a novel pathway for unsaturated fatty acid biosynthesis, in which the enzymes involved cannot be identified by conventional in silico methods. This pathway is critical to parasite survival in the host cell, since the composition of unsaturated fatty acid impacts membrane fluidity and the ability of the parasite to uptake nutrients. The lack of an orthologous pathway in the host organism provides an additional source of interest from a therapeutic perspective.

(T-PhD) **Advances on genome duplication distances**

Yves Gagnon, Olivier **Tremblay-Savard**, Denis Bertrand and Nadia El-Mabrouk

Département d’Informatique et Recherche Opérationnelle; Robert Cedergren Centre in Bioinformatics and Genomics, Université de Montréal, Montreal, QC, Canada

Whole genome duplication (WGD) is a spectacular evolutionary event that has the effect of simultaneously doubling all the chromosomes of a genome. Right after a WGD, the resulting genome contains a complete set of duplicated chromosomes. However, this initial perfect duplicate status is obscured by subsequent rearrangement events and gene losses, eventually leading to an extant rearranged duplicated genome (RD genome) containing exactly two copies of each gene, or a rearranged duplicated genome with losses (RDL genome), containing at most two copies of each gene.

Given a phylogenetic tree involving whole genome duplication events, we contribute to solving the problem of computing the rearrangement distance on a branch of a tree linking a duplication node 'd' to a speciation node or a leaf 's'. In the case of a genome 'G' at 's' containing exactly two copies of each gene, the genome halving problem is to find a perfectly duplicated genome 'D' at 'd' minimizing the rearrangement distance with 'G'. We generalized the existing exact linear-time algorithm for genome halving to the case of a genome 'G' with missing gene copies. In the case of a known ancestral duplicated genome 'D', we developed a greedy approach for computing the distance between 'G' and 'D' that is shown time-efficient and very accurate for both the rearrangement and DCJ distances.

(T-MSc) **Removing falsely predicted tRNA genes from eukaryotic species by RNA phylogenetics**

Anna **van Weringh** and Xuhua Xia

Department of Biology, University of Ottawa, Ottawa, ON, Canada

The high numbers of transfer RNA (tRNA) genes in eukaryotic organisms make the identification of organisms' gene content by experimental methods unfeasible, and computational methods essential. Though prediction methods have shown great success in identifying known tRNA genes, large numbers of false predictions are made due to the presence of repetitive elements derived from tRNA sequences. These false positives litter the predicted numbers of tRNA families, obscuring the true tRNA gene content of higher organisms. tRNA genes are known to evolve at a very slow rate, therefore a phylogenetics approach was taken to filter predicted tRNA genes. Known repetitive sequences were shown to be distinguishable by their evolutionary rate and clustering using phylogenetic analysis that accounted for both sequence and structural elements, to appropriately model the evolution of these RNA sequences.

Single tRNA genes can decode multiple codons by flexible wobble base pairing. The eukaryotic decoding pattern has previously been described as parsimonious and anticodon-saving; particular tRNA genes are avoided, specifically either a wobble G or a modified wobble A is used, but never both. It was tested if this pattern is maintained across 16 multicellular organisms by examining predicted genes that deviate from this pattern by phylogenetics. Analysis indicated that the majority of deviating wobble A and wobble G genes display long branches and clustering as observed for known false positives, with the notable exception of two wobble G families in rice.

Posters Abstracts

1. (P-PhD) **Computational DNA motif discovery for ChIP-chip and ChIP-sequencing data**

Malika **Aid**

Institute for Research in Immunology and Cancer, Université of Montréal, Montreal, QC, Canada

Transcription factors play an important role in various biological processes such as differentiation, cell cycle progression and tumorogenesis. They regulate gene transcription by binding to specific DNA sequences (TFBS). Identifying these cis-regulatory elements is a crucial step to understand gene regulatory networks. The recent developments in genomic technologies such as DNA microarrays, Chromatin immuno-precipitation followed by microarray hybridization (ChIP-chip) and ChIP-sequencing experiments have made possible the whole genome characterization of transcription factor binding sites and allow the development of several computational DNA motif discovery tools.

Although these various tools are widely used and have been successful at discovering novel motifs, they do not consider TFBS distribution properties in ChIP-chip and ChIP-sequencing data. The main drawback of these approaches is that most of the predicted motifs represent artifacts due to an inefficient assessment of their enrichment in ChIP-chip/ChIP-Sequencing data.

In this project we implemented new scoring functions which measure how DNA motifs are distributed across the ChIP fragments and how they target the set of the ChIP fragments. We showed that the implementation of these scoring functions significantly enhance the performance of DNA motif discovery algorithms by reducing the rate of false positive and false negative predictions.

2. (P-MSc) **Selective inhibition of R67 dihydrofolate reductase which is involved in bacterial resistance to trimethoprim**

Dominic **Bastien**[2,3], Natalia Kadnikova[1,2], Vincent Gauchot[1], Delphine Forge[4], Jean Jacques Vanden Eynde[4], Andreea R. Schmitzer[1] and Joelle N. Pelletier[1-3]

[1]Département de chimie, Université de Montréal, Montreal, QC, Canada
[2]PROTEO, The Quebec Network for Research on Protein Function, Structure, and Engineering
[3]Département de biochimie, Université de Montréal, Montreal, QC, Canada
[4]Laboratoire de chimie organique, Université de Mons, Belgium

Dihydrofolates réductases (DHFR) catalyse the reduction of dihydrofolate (DHF) to tetrahydrofolate (THF) which is used by cells in the synthesis of DNA. This make DHFRs an interesting target in cancer and bacterial infection treatment. The antibiotic trimethoprim (TMP) is an inhibitor of the bacterial choromosomal DHFR. TMP is commonly used to treat urinary tract infections, gonorrhoea and typhus for nearly 40 years. Today, bacterial resistance to TMP is becoming problematic. To resist to TMP, some bacterial strains express a plasmidic DHFR: the R67 DHFR. R67 DHFR catalyses the reduction of DHF to THF but is not inhibited by TMP because it's structure is totally different from the chromosomal DHFR. It's plasmidic aspect provides the means to spread this TMP resistance, making R67 DHFR a major player in increasing TMP resistance. To the best of our knowledge, no selective inhibitor (that it will not inhibit the human DHFR) is known. Our goal is to develop a specific R67 DHFR inhibitor. This inhibitor would have the potential to break the TMP restore the full effectiveness of TMP as an antibacterial agent.

Firstly, we conducted a rational screen on small hydrophobic and aromatic molecules which shared similarity with the pterin ring, which is present on DHF, by monitoring the R67 DHFR activity in presence of these compounds. We needed a co-solvant to solubilise these hydrophobics compounds, while providing conditions where R67 DHFR would be active. We thus monitored the activity of R67 DHFR in several organic solvants and ionic liquids. We determined conditions where we could solubilise the compounds while preserving good enzyme activity. We discovered nine weak inhibitors (milimolar range). Seven of these were specific to R67 DHFR. This shows that hydrophobic and aromatic characteristics can binding of a ligand on DHFR R67.

In the goal to identify more highly specifics inhibitors against R67 DHFR, we tested more complex compounds that shared some similarities with the weak inhibitors. We identified four stronger inhibitors (micromolar range). To provide insight into their potential binding mode to R67 DHFR, we performed a modeling study by molecular docking with Moldock and Autodock Vina softwares. These results suggest that intramolecular stacking interactions may dominate the binding of these compounds in the active site of the DHFR R67.

To maximize the potentiel of identifying "druggable" leads, we will enlarge our specific R67 DHFR inhibitor library. We used molecular docking with the Moldock software to screen the Maybridge virtual library. The Maybridge library consists of 80 000 pharmacophores that obey to the Lipinski rule of five and show a good ADME (absorption, distribution, metabolism and excretion) which make them good starting candidates in the process of drug development. We first screened the 80 000 compounds with a fast and efficient algorithm. Then, to obtain more reliable data, the 350 top-ranking molecules were kept for a second round of docking with Moldock with a more powerful algorithm. To confirm the results of Moldock, the 20 top-ranked from the second round were docked with Autodock Vina. Finally, 16 compounds were manually analysed and according to the possible presence of intramoledular stacking, their structures and solubility, 10 were chosen. Activity tests will be conducted to determine if they inhibit R67 inhibitors.

3. (P) **Automated organelle genome annotation with MFannot**

Natacha **Beck** and B. Franz Lang

Robert Cedergren Centre for Bioinformatics and Genomics, Biochemistry Department, Université de Montréal, 2900 Edouard-Montpetit, Montreal, QC, H3T 1J4, Canada

With recent new pyrosequencing techniques, the rate at which complete genomes are sequenced continues to 'explode' - to an extent that only highly automated analysis tools may keep up with future genome annotation and GenBank submission. Currently close to 2,500 organelle genomes are available in GenBank; yet, they have essentially been annotated manually, because available automated software (such as Glimmer, Genewise, Exonerate) does not appropriately handle organelle sequence features. The only dedicated organelle genome annotator (DOGMA; (Wyman, Jansen, and Boore 2004)) is specialized on animal mitochondrial DNAs, does not precisely assign starts and ends of genes and introns, systematically misses small exons, and some genes for structured RNAs are not recognized. Accordingly, extensive manual expert intervention is required in conjunction with DOGMA, and annotating intron-rich genomes is in reality unworkable.

For that reason we have developed MFannot. It is adapted to recognizing all specific features of mitochondrial and plastid genomes, including the two types of introns (group I and II) with distinct RNA secondary structures, structured RNAs (rRNAs, tRNAs, RNase P RNAs), and numerous protein coding genes that are highly derived in organelles and most difficult to identify (including homing endonuclease, see poster by Patrick O'Reilly). Because simple similarity searches with Blast or Fasta fail to identify most of these mentioned genetic elements, we use search algorithms such as HMMER3 ((Eddy 1998); http://hmmer.janelia.org) and Erpin/RNAweasel (Gautheret and Lambert 2001; Lang, Laforest, and Burger 2007) that excel due to the use of Hidden Markov Models (HMM) or weighted sequence profiles. In addition, Erpin/RNAweasel use specific organelle RNA secondary structure models for predictions, which further enhances sensitivity by several orders of magnitude - versus searches at the primary sequence level alone. Introns are initially predicted by combing results from RNAweasel and Exonerate (Slater and Birney 2005), and further refined by HMM procedures that allow precise delineation of intron/exon boundaries and the identification of small introns that are not found by Exonerate.
The final genome annotation is in Masterfile (Lang, Littlejohn, and Burger 1995) or in ASN format, the latter for direct GenBank submission (with only marginal manual intervention). MFannot can be accessed at http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl.

References

Eddy, S. R. 1998. Profile hidden Markov models. Bioinformatics 14:755-763.
Gautheret, D., and A. Lambert. 2001. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. J Mol Biol 313:1003-1011.
Lang, B. F., M. J. Laforest, and G. Burger. 2007. Mitochondrial introns: a critical view. Trends Genet 23:119-125.
Lang, B. F., T. Littlejohn, and G. Burger. 1995. OGMP masterfile format. URL- http://megasun.bch.umontreal.ca/ogmp/masterfile/intro.html.
Slater, G. S., and E. Birney. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.
Wyman, S. K., R. K. Jansen, and J. L. Boore. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252-3255.

4. (P-PhD) **Inferring and validating horizontal gene transfer events**

Alix **Boc**

Université du Québec à Montréal, Montreal, QC, Canada

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.

5. (P-PhD) **A hierarchical model for evolution of ribosomal RNA**

Sergei V. Chteinberg and Konstantin **Bokov**

Biochemistry Department, Université de Montréal, Montreal, QC, Canada

Analysis of the tertiary structure of the 23S ribosomal RNA allowed us to determine the order in which different rRNA elements were added to the ribosome as it evolved. The analysis was based on the following suggestions: (1) each new element emerged as a single insertion into the rRNA polynucleotide chain; (2) in each double-helical region, both strands emerged simultaneously as parts of the same element; (3) when a double helix and an adenosine stack formed an A-minor interaction, the element containing the double-helix was considered a more ancient acquisition of the ribosome than the element containing the adenosine stack.

Based on these suggestions, we developed an iterative procedure of gradual dismantling the tertiary structure of the ribosomal RNA through removal of those elements that were considered more recent acquisitions of the ribosome. After 59 acts of removal, the remaining element was 220 nucleotide-long and corresponded to the central part of domain V. This element was thus considered the initial one, from which the evolution of the 23S rRNA commenced. All other elements were gradually added to this element as insertions containing all necessary details to dock with the surface of the evolving ribosome without disturbing already existing parts.

References

Bokov, K. and Steinberg, S.V. Nature 457, 977-980 (2009)

6. (P-MSc) **Prediction of regulatory loops involving miRNAs and genes regulated by retinoic acid receptor**

Asma **Boufaden**, Sylvie Mader

Institute for Research in Immunology and Cancer, Department of Biochemistry, Université de Montréal, QC, Canada

The retinoic acid receptor (RAR) is a type of nuclear receptor that is activated by the ligand retinoic acid (RA). The retinoid signal is transduced by two families of nuclear receptor, RARs and the retinoid X receptors (RXRs), which works as RXR/RAR heterodimers and bindd a specific DNA sequences or RA response elements (RAREs) in the promoters of a large number of retinoid-target genes. The mechanism of regulation of RAR is altered in breast cancer cell lines due to a reduced capacity to synthesize RA. Also aberrant patterns of miRNA expression have been reported in human breast cancer and a number of genes involved in breast cancer progression have been identified by in silico analysis to be targets of miRNAs that are deregulated in breast cancer. The miRNAs are small, non-protein-coding RNAs of 19-25 nucleotides that could play a role in the mechanism of regulation of RAR. In fact, in the literature some regulatory loops involving the estrogen receptor α (ERα), miRNAs and ERα-target genes have been reported in human breast cancer.

We aim to predict regulatory loops between RAR and miRNAs targeting the same genes in breast cancer cell lines (MCF7 and SKBR3) treated by RA. We propose to integrate ChIP-on-chip datasets and microarrays datasets by using miR targets prediction tools to identify such regulatory loops. The regulatory loops will be then filtered, in order to reduce the number of false positive, based on databases designed to represent human miR expression profiles in different tissues or cell types. Finally, we suggest the validation of some potential regulatory loops.

7. (P-MSc) **Study of the mechanism of entry of Anthrax toxins in the cell**

Lucien **Fabre** and Isabelle Rouiller

Department for Anatomy and Cell Biology, McGill University, Montréal, QC, CANADA

The anthrax toxins are part of the AB toxin family in which B moetie binds to the cell membrane allowing subsequent translocation of A moetie. In the case of anthrax, the B moetie is represented by the protein PA (protective antigen) and the A moetie by the two proteins EF (Edema Factor) and LF (lethal factor).

After being recruited by the Capillary Morphogenesis protein 2 (CMG2) cell receptors, PA organizes itself into an heptamer form. It can bind up to three ligands (either EF and LF) before being endocytosed. Current models suggest that the decrease of pH inside the endosomes allows a conformational change of PA from pre-pore form to pore form and allows EF and LF ligands to pass through the pore to enter the cytoplasm. However, the pore diameter is about ten times lower than the diameter of the ligands (10 Å against 100 Å). A process of folding / unfolding has been proposed but remains controversial. Indeed, LF is active even if it is attached to PA63 to which they are associated with a high affinity constant.

To identify the actual process of transition factors EF and LF into the cytoplasm, we propose to determine by cryo-electron microscopy combined with image analysis of three-dimensional structures of complexes formed by PA and LF at different stages of entry into the cell. We have started by determining the 3D reconstruction of the PA-LF complex in a the pre-pore conformation using 3D cryo-EM. Although, our map shows that three LF molecules can bind on top of the PA heptamer, it is clear that the three LF molecules are not in the same conformation. Coupled with molecular dynamics study, our study points to a mechanism leading to the start of translocation.

8. (P-PhD) **Synergistic growth inhibition by retinoic acid and Herceptin in HER2/RARA co-amplified breast cancer cells**

Marieke Rozendaal, Slim **Fourati** and Sylvie Mader

Biochemistry Department, Université de Montréal, Montreal, QC, Canada

Herceptin, a humanized monoclonal antibody that inhibits the tumorigenic effects of HER2 is currently used as a common treatment strategy for HER2 positive breast cancer. Success rates of Herceptin in the clinic are good; however a majority of patients that originally respond will develop resistance within a year.

We have observed that in a breast cancer cell line that carries a co-amplification of the HER2 and RARA genes retinoic acid (RA) and Herceptin function in a synergistic way to inhibit proliferation. Thus, lower doses of both drugs suffice to obtain similar anti-tumor activity. We have identified several candidate genes that may mediate the synergy. Since 1/6 of all tumors carry a co-amplification of the HER2 and RARA genes, we propose this subgroup of tumors could benefit from co-treatment with RA and Herceptin.

9. (P-PhD) **The central role of the reverse-Hoogsteen base-pair U54-A58 in the folding of the DT region**

Tetsu M. **Ishii** and Sergei Chteinberg

Biochemistry Department, Université de Montréal, Montreal, QC, Canada

To understand the mechanism of tRNA folding, we used a molecular modeling approach in the analysis of the role played by the reverse-Hoogsteen base pair U54-A58 in the structuring of the DT-region. We used the set of all T-loop-like structures available in the pdb-database. The analysis revealed a chain of structural events linked through cause-effect relationships that leads to the formation of the standard conformation of the DT region. In the available RNA tertiary structures there are several dozen cases where U and A form a reverse-Hoogsteen base pair.

Surprisingly, A never is found stacked to its 5'- neighbor. Molecular modeling shows that such stacking is prohibited due to the interference between the backbone connecting the A with its 5'-neighbor and the O6 atom of the U. Because of this interference, the 5-neighbor of the A can either stack to the U or distance itself from the A, opening the space between the two nucleotides. Both types of arrangements are found on many occasions in the known RNA structures. However, if the U and A are separated in the polynucleotide chain by less than five nucleotides, as it happens in the tRNA structure, the 5'-neighbor of the A cannot stack to the U and is thus destined to stay at a double distance from the A. Such juxtaposition of the A and of its 5'-neighbor can be stabilized by the intercalation of an additional nucleotide between them (G18 in the standard tRNA structure). Still, there are known cases where the two nucleotides remain separated without an intervening nucleotide. This indicates that the intercalation can only stabilize the opening of the space between the A and its 5'-neighbor, but does not cause it.

More importantly, we show that when the reverse-Hoogsteen bp U54-A58 is directly stacked to a WC bp 53-61, there is an interference by the H1 atom of N61 with A58. This interference however is avoided by the introduction of a bulge between A58 and N61. If a pyrimidine occupies N61, due to its bulky 6-member ring, the proper positioning of N61 and the A would require that the bulge between them contain at least two nucleotides. Such situation happens in the tRNA, where C61 is separated from A58 by the bulge 59-60. However, if N61 is a purine, its smaller 5-member ring would allow the bulge to consist of only one nucleotide, as it happens in many known T-loop-like structures outside tRNA. Thus, the presence of the reverse-Hoogsteen base-pair U54-A58 ensures the reliable folding of all elements that compose the T-loop and can explain its high conservation among cytosolic tRNAs.

10. (P-MSc) **Next-generation RNA sequencing (RNA-Seq) technology allows for characterization of the whole transcriptome at an unprecedented single-base resolution - CANCELLED**

Emili **Lalonde**

McGill University Montreal, Montreal, QC, Canada

On average, 95 million short read sequences (~36-76 bp) are being generated per sample, presenting significant computational challenges and requiring novel strategies to fully and accurately extract this information from the data. In particular, RNA-Seq is able to reveal point mutation and short indels, gene fusion transcripts arising from genomic rearrangements, and novel transcript isoform variants arising from aberrant alternative splicing, all of which are known to contribute to cancer pathogenesis. Here, we use RNA-Seq (Illumina Genome Analyzer II) to characterize the transcriptomes of three primary human breast tumours and four breast cancer cell lines (HCC1937, HCC3153, SUM149 and SUM131502), all of which harbour BRCA1 mutations. We present examples of commonly mutated genes, gene fusions and recurrent splicing events identified in these samples with both novel and published strategies.

The mutational spectrum of these BRCA1 breast cancer samples has been extensively analyzed. We have found several genes with either homozygous mutations or compound heterozygous mutations in several of these samples. Additionally, we have identified three gene fusions from our analysis, including two novel fusions. We also examined genes involved in aberrant alternative splicing events in the BRCA1 samples compared to a broader group of breast cancer samples. We will show that the genes affected in the two groups differ in important functional pathways. Finally, we identified a number of overexpressed aberrant alternative splicing events in the BRCA1 samples which likely have important functional consequences in terms of gene function. For example, alternative splicing events were identified where exons coding for important protein domains are spliced out, or, where skipped exons and introduce frameshifts and premature termination codons.

This work will help describe the transcriptomic landscape of BRCA1-deficient breast cancer samples. Interactions between these novel lesions and known BRCA1 mutations may also identify key pathways contributing to the observed phenotype. These and similar findings will continue to become more valuable as we transition away from single-marker targeted treatment to a systems biology-based intervention.

---

11. (P-PhD) **Modeling contaminants in TAP-MS/MS experiments**

Mathieu **Lavallée-Adam** et al. → see Talks

---

12. (P) **Characterization of a genetic determinant of self-incompatibility in the micro-endemic plant *Biscutella neustriaca* (Brassicaceae) and implications for its conservation**

Jean-Baptiste **Leducq**[1], Célia Gosset², Matthieu Poiret², Vincent Castric², Sylvain Billiard² and Xavier Vekemans²

[1] Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, QC, Canada
² Laboratoire de Génétique et Evolution des Populations Végétales (GEPV), UMR 8016, CNRS, Université Lille 1,Villeneuve d'Ascq, France

Self-incompatibility (SI) has evolved in many plant families to allow selfing avoidance and to limit inbreeding depression. SI is controlled by a highly polymorphic locus (S-locus), with dominance relationships that can modulate allelic expression in the case of sporophytic SI (SSI), as found in Brassicaceae. In this family, the S-locus involves two main genes encoding for proteins respectively in pollen (SCR) and pistil (SRK) which are responsible of self-pollen recognition leading to its rejection. Here, we focused on the SRK gene in order to characterize allelic richness at the S-locus in the micro-endemic plant *Biscutella neustriaca*.

Thanks to controlled pollinations, we determined that *B. neustriaca* was strongly self-incompatible. By using TA-cloning and PCR amplifications, we identified 45 SRK-like sequences belonging to two distinct phylogenetic groups, with 1-4 sequences present by individual, some sequences of the same group always segregating by pairs. Thanks to controlled crosses, we validated the linkage of these sequences or sequence pairs with the expression of at least 12 specificities at the S-locus (S-alleles), with dominance relationships modulating their expression as previously found in other Brassicaceae species. Additionally, we determined that two main dominance levels corresponded to both phylogenetic groups of S-alleles, as found in the *Brassica* genus, although both groups in *B. neustriaca* were phylogenetically distinct to those of Brassica. We suggest that each identified sequence pair corresponds to a functional SRK copy linked with either a non-functional copy of SRK or another gene belonging to the S-locus, confirming the complexity of the S-locus architecture previously highlighted in *Brassicaceae*.

In small plant populations, low allelic diversity at the S-locus is expected, and this could affect female reproductive success by decreasing the proportion of compatible mates, thereby generating a S-locus specific Allee effect. From our genetic data, we estimated diversity at the S-locus in experimental and natural populations of *B. neustriaca* to investigate such effect on maternal reproductive success. We discuss implications of the S-locus specific allele effect in conservation of endangered SI plants.

13. (P-PhD) **Armadillo v1.0 - A graphical-based platform for phylogenetic simulations**

Etienne **Lord**, Mickael Leclercq, Abdoulaye Baniré Diallo, Vladimir Makarenkov.

Département d'informatique, Université du Québec à Montréal, Montreal, QC, Canada

Phylogenetic analysis is one of the cores of biological research and is a necessary task for researcher studying horizontal gene transfer events (HGT) and ancestral sequence reconstruction. However, while computer scientists can develop their own computer scripts to perform multiple simulations, traditional biologists could find it difficult to interface multiple bioinformatics algorithms and applications into a data pipeline.

We report here the extension of a workflow platform, which was originally used as an educational tool to help students understand the basic phylogenetic inference process using the PHYLIP package. This new framework, developed in Java, can now help life science researchers rapidly develop data-flux prototypes or simulations with automatic multiple species phylogenies determination. The current version include custom interfaces for commonly used phylogenetic applications such as search for orthologous genes, multiple sequence alignment (MSA), evolution model determination, inference of phylogeny using maximum likelihood or maximum parsimony, simulation of sequence data and tree metrics. This platform also implements persistence of those in silico experiments by recording methodologies, computational pipelines and results inside a single project file, allowing easy sharing of results between researchers. The platform and supporting tutorials can be found at: http://adn.bioinfo.uqam.ca/armadillo.

14. (P-PhD) **Trans-splicing mediated by guide-RNA?**

Seyed Amir **Malekpour**[1], Marcel Turcotte[2] and Gertraud Burger[1]

[1] Robert Cedergren Centre in Bioinformatics and Genomics, Biochemistry Department, Université de Montréal, Montreal, QC, Canada
[2] School of Information Technology and Engineering, University of Ottawa, Ottawa, ON Canada

Recently, highly fragmented genes have been discovered in the mitochondrion of the single-celled eukaryote *Diplonema papillatum*. In this organism, each gene piece (module) resides on a distinct mitochondrial chromosome. It was demonstrated that gene modules are concatenated at the RNA level via trans-splicing, but the mechanism by which modules recognize each other is unknown. None of the consensus sequence elements typical for spliceosomal, group I and group II introns were detected adjacent to coding regions. Further, the *cox1* transcript is edited by addition of a run of six nonencoded uridines inserted between two gene modules.

We assume that trans-splicing of gene modules is mediated by small RNAs. This hypothesis was inspired by a phenomenon observed in mitochondria of kinetoplastids, the sister group of diplonemids, where guide RNAs (gRNAs) direct RNA editing. In kinetoplastids, editing consists of uridine insertions and deletions and is very similar to editing of the *Diplonema* *cox1* transcript. We assume that in *Diplonema*, both concatenation of fragment transcripts and editing is mediated by gRNAs.

The sequence of kinetoplastid gRNAs is complementary to the stretches of pre-mRNA around editing site. To detect potential gRNAs in *Diplonema* we used regular expressions to search sequence motifs that are complementary to cognate module transcript junctions. For the eight junctions of *cox1*, we found numerous potential gRNAs. Since gRNAs are expected to have shared properties we classified them based on several criteria, (i) common gRNAs bridge length (unaligned bases of gRNAs vis-à-vis of the junction), (ii) distance of match from the junction, and (iii) the source of gRNAs (mitochondrial or nuclear genome). Further, non-unique gRNAs are eliminated from these classes. Uniqueness implies that a given gRNA is not able to direct trans-splicing of nonconsecutive modules (mis-joining).

We will present which classes of predicted gRNAs serve a maximum number of junctions. Biologically significant classes likely include those with a minimum portion of nucleotides involved in GU pairs and, according to kinetoplastid gRNAs, a minimum bridge length and a maximum number of matches close to junctions.

15. (P) **Bringing order to disorder: genomic analysis uncovers three distinct forms of protein disorder**

Magali **Michaut** et al. → see Talks

16. (P-PhD) **Nuclear gene structure of a highly divergent protozoa**

Sandrine **Moreira**[1], Marcel Turcotte[2] and Gertraud Burger[1]

[1] Robert Cedergren Centre in Bioinformatics and Genomics, Biochemistry Department, Université de Montréal, Montreal, QC, Canada
[2] School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada

Studies in our laboratory are focused on novel gene expression mechanisms and genome structures. The organisms of choice are a group of poorly investigated unicellular eukaryotes, the diplonemids (Euglenozoa). Euglenozoa are thought to be one of the earliest diverging eukaryotic groups, much before the emergence of the well-studied animals, fungi and plants.

The mitochondrial genome of diplonemids is highly unusual. First, this genome is composed of a hundred or so circular chromosomes. Further, all genes encoded by the mitochondrial DNA are split into several pieces (modules) each of which is located on a different chromosome. In addition, gene modules are transcribed separately into RNA and then joined to a complete messenger RNA by trans-splicing. Finally, at least one mitochondrial pre-mRNA is modified in sequence post-transcriptionally, and this RNA editing proceeds by addition of uridines exactly at the junction of two modules (Marande & Burger 2007; Vlcek et al. 2010; Kiethega et al, unpublished). Uridine-based RNA editing is known from the diplonemids sister group, the kinetoplastids (addition and deletion of bases), but this process being interlinked with trans-splicing is unheard of.

RNA editing and trans-splicing are likely performed by a multifunctional molecular machine. To identify the genes involved in this machinery, a project for sequencing the nuclear genome of the diplonemid species *Diplonema papillatum* was initiated, using the 454 massive parallel pyrosequencing technology of Roche Life Science. This platform produces reads of about 300 bases length. A rough sequence assembly with the MIRA assembler of the first set of reads available led to 1 Mb of contigs (around 2% of the total length of the genome). Moreover, a library of about 4000 EST clusters has previously been generated in the framework of the pan-Canadian Protist EST Program (Keeling et al 2005).

The first questions we asked aims at the structure of nuclear genes, (i) are they discontinuous as the mitochondrial genes of Diplonema or rather orthodoxically contiguous? (ii) Do they contain introns, and if yes, how many and which type? For that, we mapped EST clusters on the genomic contigs using the Exonerate software.

As a preliminary result, there is no indication of an unusual gene structure such as fragmented coding regions. Further, we observed that at least 90 % of the genes are intron-less, which will greatly facilitate genome annotation. We will report about the set of genes detected so far, intron splicing rules, and size variation of introns, in addition to read statistics and genome coverage. Since this first set of 454 reads provides obviously limited genome coverage and thus a relatively low sequence quality, we will need to validate our preliminary findings as soon as new data sets come in (esp. paired end sequences).

17. (P-PhD) **The correlation between gene length and nucleotide content: a comparative study**

Hamid **Nikbakht**, Mihai Albu, Donal A. Hickey

Department of Biology, Concordia University, Montreal, QC, Canada

GC content is one of the simplest yet most important attributes of a genome which can be used for a better understanding of the basic makeup of a genome and the evolution of the coding sequences. The broad variation in nucleotide content among different genomes of different species as well as within a single genome of one species has been studied by different groups. There are several theories explaining this variation although none of them can inclusively explain both the variation between different genomes as well as within a single genome.

In this study we have described a correlation between gene length and the nucleotide content and explained how the length of a gene can affect its content. We have analyzed the low GC genome of honeybee and compared it with rice genome which is a GC rich genome. We have compared the average length of coding sequences and introns as well as total genes with their GC content. We also have studied the effect of the GC content of the genes on the amino acid content of their corresponding proteins.

Our results show a negative correlation between the nucleotide bias and the average length on coding sequences as well as introns and total gene. We also show that nucleotide content of a gene affects the amino acid content of its corresponding protein. Our study shows that shorter genes on average are more biased comparing the longer ones. Considering a constant rate of both neutral and harmful mutations, we explain that longer genes provide larger and more accessible targets for the harmful mutations. This makes them more likely to be eliminated from population and holds them back from being affected by different evolutionary forces, such as biased mutations, as fast as shorter ones. This idea resembles the background selection idea but in a finer scale of a single nucleotide. We also show that the bias in nucleotide content affects the content of the amino acids of the corresponding proteins, but this effect is not due to the effect of the bias on the nucleotide content of the codons.

18. (P-MSc) **Finding and classifying homing endonucleases - an HMM approach**

Patrick **O'Reilly** and B. Franz Lang

Robert Cedergren Centre in Bioinformatics and Genomics; Biochemistry Department, Université de Montréal, Montreal, QC, Canada

Homing endonucleases cleave genomic (double-stranded) DNA at highly specific sequence motifs of ~20 bp in length. They are most frequently encoded within organelle introns, but also occur as freestanding open reading frames in organelle genomes and within inteins (protein introns). Homing endonucleases permit intron transposition, by cleaving intron-less genes precisely at their corresponding intron insertion site, followed by recombinative repair with the intron-containing gene version (Dujon et al. 1986). In the literature, four families of homing endonucleases are defined based on short, conserved sequence motifs: LAGLIDADG, GIY-YIG, H-N-H and HIS-CYS box (Belfort and Roberts 1997). To test the validity of this classification with a formal bioinformatics approach, we built profile Hidden Markov Models (HMM; (Eddy 1998)) with a comprehensive collection of organelle intron proteins that were downloaded from the GOBASE database (O'Brien et al. 2009). In a first round of analysis, we used the JackHMMer program (http://hmmer.janelia.org) for iteratively searching all individual sequences against the complete data collection, building and refining Markov models at each step. We thus obtain 4572 groups (equal to the initial number of proteins), from which we built a matrix to identify overlaps among groups, thus reducing their number to ten. We find that the given protein annotations in three groups is perfectly consistent with three of the four described protein families: LAGLIDADG, GIY-YIG and H-N-H. Intron proteins within the other seven groups have various other activities (such as mito-ribosomal proteins), but no recognized endonuclease activity.

The next step was building comprehensive Markov models for the three endonuclease groups with HMMFinder (see poster by J-F. Théroux), which automates the task of multiple sequence alignment with Muscle (Edgar 2004), followed by iterative model building and refinement with Hmmer. We find that the three endonuclease HMM models identify all proteins of its class, with high statistical support (> E-4). When applying these models in the analysis of mitochondrial DNAs of three mycorrhizal fungi (Glomus irregulare 494 and PR, and Glomus diaphanum), we find that several previously unassigned open reading frames are in fact members of the LAGLIDADG and GIY-YIG endonuclease families. They appear to constitute non-intronic mobile elements, similar to those previously characterized in the fungus Allomyces macrogynus (Paquin, Laforest, and Lang 1994).

References

Belfort, M., and R. J. Roberts. 1997. Homing endonucleases: keeping the house in order. Nucleic Acids Res 25:3379-3388.

Dujon, B., L. Colleaux, A. Jacquier, F. Michel, and C. Monteilhet. 1986. Mitochondrial introns as mobile genetic elements: the role of intron-encoded proteins. Basic Life Sci 40:5-27.

Eddy, S. R. 1998. Profile hidden Markov models. Bioinformatics 14:755-763.

Edgar, R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

O'Brien, E. A., Y. Zhang, E. Wang, V. Marie, W. Badejoko, B. F. Lang, and G. Burger. 2009. GOBASE: an organelle genome database. Nucleic Acids Res 37:D946-950.

Paquin, B., M. J. Laforest, and B. F. Lang. 1994. Interspecific transfer of mitochondrial genes in fungi and creation of a homologous hybrid gene. Proc Natl Acad Sci U S A 91:11807-11810.

19. (P-PhD) **Differential selection profiles using phylogenetic models for understanding HIV adaptation**

Sahar **Parto** and Nicolas Lartillot

Robert Cedergren Centre in Bioinformatics and Genomics, Department of Biochemistry, Université de Montréal, QC, Canada

AIDS, which is one of the most challenging current diseases, does not have any cure or vaccine yet. The extensive rate of HIV mutation and adaptation makes the design of vaccine difficult, as it enables the virus to escape from the immune system (escape mutation). So the first step for designing efficient vaccine is to identify consistent patterns in viral adaptation, as a function of the specific genetic background of the host. It has been shown that polymorphisms in HIV-I are associated with particular host HLA (Human Leukocyte Antigen) alleles. This association confirms the effect of HLA restricted CTL (Cytotoxic T Lymphocyte) response on HIV evolution.

In this study, a differential mutation-selection model is developed which parametrizes mutational and selective effects bearing on the overall substitution process. it is implemented in a Bayesian MCMC framework which allows us to tease out each component of the model and estimate differential selection profiles; one distinct selection profile is estimated for each host genetic background and specifies which amino acids are selected for or selected against at each position of the viral coding sequences. This model is used to analyze the data of patients with identified genetic immune profile.

Results show that some amino acids are selected differently in specific positions of HIV sequence in patients with different HLAs. By associating specific viral adaptation with specific host genetic background, it is possible to understand how HIV escapes from immune system, which in turn provides useful guideline to design an efficient vaccine against AIDS.

Keywords: selection, escape mutation, virus adaptation, Bayesian analysis, phylogenetic tree, MCMC, HLA, MHC

20. (P-MSc) Adaptive selection of HIV in HLA epitopes is associated with ethnicity in Canada

Manon **Ragonnet-Cronin**[1,2], S. Aris-Brosou[2], I. Joanisse[1], H. Merks[1], D. Vallee[1], K. Caminiti[1], M. Rekart[3], M. Gilbert[3], P. Sandstrom[1] and J. Brooks[1]

[1] National HIV and Retrovirology Laboratories, Public Health Agency of Canada, Ottawa, ON, Canada
[2] Department of Biology, University of Ottawa, Ottawa, ON, Canada
[3] BC Centre for Disease Control, Vancouver, BC, Canada

Host immune selection pressure favours the development of mutations in HIV that allow immune escape. However, it is unknown if immune pressure among an HIV infected population that share similar human leukocyte antigens (HLA) will lead to convergent evolution in HIV. As ethnic groups have distinct and characteristic HLA allele frequencies, can we expect to observe convergent viral evolution within ethnicities, irrespective of viral history?

Here, we address this question by sequencing and analyzing the pol gene sampled from 1481 individuals living in BC, Canada. No phylogenetic pattern of ethnicity could be detected. Yet, in terms of Fst values, pol sequences showed significant differentiation between ethnicities, both at the nucleotide and amino acid level. A total of 22 amino acid sites were inferred to be under positive selection across all ethnic groups, a third of which showed a fixation pattern that is ethnic-specific. As these latter sites are tightly associated with HLA in HIV Subtype B, our results suggest that escape mutants are, in part, shaped by host ethnicity. The implications of these findings, with respect to treatment and vaccine development, are discussed.

21. (P-PhD) **Markov Chain Monte Carlo computational analysis of chromosome conformation capture Carbon copy data**

Mathieu **Rousseau** et al. → see Talks

22. (P-PhD) **HyperGenometrics: hypergeometric-based analysis of genomics data for prediction of molecular functions and biological processes**

Hamed **Shateri-Najafabadi**[1,2] and Reza Salavati[1,2,3]

[1] Institute of Parasitology, McGill University, McDonald Campus, Montreal, QC, Canada
[2] McGill Centre for Bioinformatics, Montreal, QC, Canada
[3] Department of Biochemistry, McGill University, Montreal, QC Canada

Protein molecular functions are widely predicted from protein sequences based on presence of protein domains and/or certain protein folds. This is while the information about short linear motifs that are associated with molecular functions is often disregarded. Here, we describe a pipeline that finds short protein motifs that are over-represented within different protein categories, such as different molecular functions, and then combines these short motifs using a naïve Bayesian network in order to predict other proteins that are associated with those categories. Applying this pipeline, which is called HyperGenometrics, to the entire set of proteins with experimentally verified molecular function annotations in GO database, we have found several thousand short motifs that are significantly associated with different molecular functions. These short motifs are often similar to known sequences of active/binding sites of proteins. Furthermore, many molecular functions can be effectively predicted based on these motifs. We have used this method to predict molecular functions for several previously uncharacterized proteins of the human parasite Trypanosoma brucei. Our results suggest that short protein sequences not only can correctly predict the molecular functions of characterized T. brucei proteins, but also can make new predictions for previously uncharacterized proteins. Furthermore, we have applied this method to T. brucei protein categories that, instead of GO molecular functions, are based on KEGG pathway annotations. HyperGenometrics finds hundreds of short motifs that are associated with different pathways. These motifs may either represent the active sites of the most important and prevalent enzymatic reactions of each pathway, or may represent pathway-specific sites for post-translational regulation of proteins. Using the HyperGenometrics pipeline, we have been able to combine these protein motifs with other genomic features in order to predict several pathways to which previously uncharacterized T. brucei proteins belong.

HyperGenometrics is available online at http://webpages.mcgill.ca/staff/Group2/rsalav/web/software.htm.

23. (P-PhD) **Using A quantitative structure-activity approach to design functional RNA sequences**

Karine **St-Onge**[1,2], Sylvie Hamel[1], and François Major[1,2]

1 Computer Science Department, Université de Montréal, Montreal, QC, Canada
2 Institute for Research in Immunology and Cancer, Robert Cedergren Centre in Bioinformatics and Genomics, Université of Montréal, Montreal, QC, Canada

RNA function is due to the presence on the surface of its 3D structure of specific and precisely positioned chemical groups. Predicting that an RNA sequence will exhibit or not a given function consists in predicting that it will fold so to expose these chemical groups.

Here, we developed an RNA 3D QSAR method to make such predictions, and applied it to the sarcin-ricin loop (SRL). We used twelve SRL sequences, 3D structure prediction, electrostatic profiling, and a support-vector machine (SVM) to derive the structural profile of functional SRLs and a computational classifier that can predict the viability of new SRL sequences.

The derived profile matches more than 98% among 806 species of the SRL sequences in a ribosomal RNA alignment. We can now use the classifier to predict the viability of new SRL sequences.

24. (P-PhD) **Thermodynamic of Z-pro-proline binding pathways to porcine prolyl oligopeptidase**

Jean-François **St-Pierre**[1], Alex Bunker[2], Normand Mousseau[1]

[1] Département de Physique, Robert Cedergren Centre in Bioinformatics and Genomics, Université de Montréal, Montreal, QC, Canada
[2] Center for Drug Research, Helsingin Yliopisto, Finland

The Prolyl oligopeptidase (POP) family of proteins is a group of endopeptidase able to cleave peptides of up to 30 amino acids on the C-terminal side of internal prolines. While its role is not entirely clear, we know that POP has many neuroactive peptide substrates like substance P, beta-endorphine, neurotensine and oxytocine. It has been show that inhibition of POP can re-establish chemically induced memory loss and it is a target for treatment of many cognitive disorders.

The protein structure of porcine POP has been elucidated and is composed of two domains, a catalytic alpha/beta fold domain covering a beta-propeller domain. The co-crystallized inhibitor z-pro-prolinal (ZPP) is found between these 2 domains linked to serine 554 by a hemiacetal reversible bond. In this work, we are interested in the pathway taken by ZPP to access this inner cavity and bind to POP. We examine two pathways, through a flexible loop located at the interface of the two domains of POP and through the cavity in the center of the beta-propeller domain.

Two methods have been used to thermodynamically sample the entry and exit pathways of ZPP. First, steered molecular dynamics by which ZPP was tied to a virtual spring and pulled out of POP in a molecular dynamics (MD) simulation. Second, using umbrella sampling where many positions of ZPP along the pathway where launched in individual MD simulations with their position relative to POP restrained by a spring. Both methods measure the tension on the spring and can convert this force profile into a free energy profile along the exit pathways from which the difference in the transition energy peaks indicates what pathway is the most probable. Furthermore, detailed analysis of the exit pathway allow to identify amino acids who may play a crucial role in recruiting the inhibitor.

25. (P-MSc) **Identifying proteins in large genomic datasets with profile HMM models**

Jean-François **Théroux** and B. Franz Lang

Robert Cedergren Centre in Bioinformatics and Genomics; Biochemistry Department, Université de Montréal, Montreal, QC, Canada

Genes are usually identified by similarity searches with tools such as Blast or Fasta, which works well with closely related sequences - but may fail for up to 50% of genes in evolutionary distant species. The sensitivity of identification increases dramatically when using as search reference sequence profiles (i.e., statistical weighting of sequence positions based on multiple sequence alignments), instead of single sequences. The goal of our project is to improve tools that allow identification of little conserved genes using Hidden Markov Models (HMMs). HMMs use statistics to describe the probability distribution of character strings against a model (Eddy 1996). Similarity search tools based on HMM profiles have not only highest sensitivity, but a recent implementation (HMMER3; (Eddy 1998); (Eddy 2009)) has execution times as fast as Blast.

Yet, HMMER3 is not easily used in analysing large genomic datasets, as it requires frequent user intervention (formatting of input data to comply with an inconveniently strict format, multiple alignments, model building, HMM searches one at a time, parsing and reformatting of results, etc.). We have therefore created a new tool (HMMfinder) that uses HMMER3 as a search engine, and that automates all time-consuming steps, providing a user-friendly alternative to HMMER3. In addition, we have implemented a normalization method that removes phylogenetically close sequences prior to building HMM models. It results in significantly more sensitive search models (often several orders of magnitude) due to the removal of bias that is introduced by groups of similar sequences.

We have subsequently tested HMMfinder for searching regulatory genes (families of transcription factor) that are little conserved among eukaryotes and that are involved in constituting multicellularity. Multicellular organisms are characterized by differentiated cells that communicate and work together. It allows organisms to delegate precise tasks to specific cell types, or ultimately tissues and organs, using a process known as cellular differentiation. The cellular properties required for multicellularity (like cellular adhesion and cellular communication) depend on specific proteins. These are well characterised in metazoans but not in other eukaryotes. We will present our results demonstrating the occurrence of transcription factor genes previously thought to be animal-specific, in protists and fungi.

References

Eddy, S. 2009. HMMER3: a new generation of sequence homology search software. http://hmmmer.janelia.org.
Eddy, S. R. 1996. Hidden Markov models. Curr Opin Struct Biol 6:361-365.
Eddy, S. R. 1998. Profile hidden Markov models. Bioinformatics 14:755-763.

26. (P-MSc) **Mapping regulatory sites using allelic imbalance data**

Emilio **Vello**[1], Jean-François Lefebvre[1], Tomi Pastinen[2,3] and Damian Labuda[1,4]

[1] Sainte-Justine Hospital Research Centre, Université de Montréal, QC, Canada
[2] McGill University and Genome Quebec Innovation Centre, Montréal, QC, Canada
[3] Department of Human Genetics and Department of Medical Genetics, McGill University, Montreal, QC, Canada
[4] Department of Pediatrics, Robert Cedergren Centre in Bioinformatics and Genomics, Université de Montréal, Montreal, QC, Canada

Wilson and King (1975) proposed that evolution frequently operates through mutations affecting genetic regulation. Likewise, it is expected that genetic variation responsible for inter-individual differences will be due to variation in regulatory sites. Identifying such sites is thus important in the genetic and medical research. Individuals carrying different regulatory alleles will exhibit allelic imbalance(AI) due to differential expression of regulated transcript from two parental copies of the same locus.

We propose two novel statistical tests to map regulatory sites using allelic imbalance data and whole genome information on single nucleotide polymorphisms (SNPs). A regulatory site that is in linkage with a SNP forms a two-site haplotype. For biallelic sites, there are only 4 possible haplotypes. In the absence of recombination, only three of such haplotypes can be observed in three distinct combinations, each combination reflecting a particular mutational history with the underlying genealogy. Each combination will lead to a different non-random distribution of the alleles of an analyzed SNP, between AI individuals representing regulatory site heterozygotes and non-AI individuals. Testing for deviation from random distribution can be done knowing chromosomal phase or directly with the genotypes.

Here we compare haplotype-based (phase known) and genotype-based test in simulated and in two sets of experimental data (Ge et al. 2009 and Montgomery et al. 2010). In addition we verified our results with independent data on selected loci available in the literature. We have computed the false discovery rate by randomly assigning individuals as showing allelic imbalance and running our tests across the whole genome a number of times which yielded a very low number of significant SNPs. We have also computed by simulation the power of the tests.

As conclusion, each method has its advantages, the haplotype based tests have a high power and the genotype based tests are not sensible to phasing error. In order to find a regulatory site in the most effective manner, both methods should be used in a complementary way.

Contacts

Colloquium participants

version Friday, November 5, 2010

Family name, given name	*Role*	*Institution*	*Courriel*
Aid Malika	P	UdeM	malika.aid@umontreal.ca
Babaï Dariouch		UdeM	babaid@umontreal.ca
Badescu Dunarel		UQAM	dunarel@gmail.com
Bastien Dominic	P	UdeM	Dominic.Bastien@Umontreal.ca
Beck Natacha	P, Volonteer	UdeM	natabeck@gmail.com
Boc Alix	P	UQAM	boc.alix@courrier.uqam.ca
Bokov Konstantin	P	UdeM	konstantin.bokov@umontreal.ca
Boufaden Asma	P	UdeM	asma.boufaden@umontreal.ca
Brinkmann Henner	Jury	UdeM	henner.brinkmann@umontreal.ca
Burger Gertraud	Organizer	UdeM	gertraud.burger@umontreal.ca
Chteinberg Serguei	Jury	UdeM	serguei.chteinberg@umontreal.ca
Csuros Miklos	Jury	UdeM	csuros@iro.umontreal.ca
Diss Guillaume		U.Laval	guillaume.diss.1@ulaval.ca
Doroftei Andrea		UdeM	doroftel@iro.umontreal.ca
Droit Arnaud	Jury	U.Laval	arnaud.droit@crchuq.ulaval.ca
El-Hachem Maud		UdeM	maud.el-hachem@umontreal.ca
El-Mabrouk Nadia	Jury	UdeM	mabrouk@iro.umontreal.ca
Fabre Lucien	P	UdeM	lucien.fabre@umontreal.ca
Filiatreault Isabelle		UdeM	isabelle.filiatreault@umontreal.ca
Foisy Sylvain	Jury	MHI	sylvain.foisy@inflammgen.org
Fourati Slim	P	UdeM	slim.fourati@umontreal.ca
Fournier Frédéric		U.Laval	fournier.frederic@gmail.com
Freschi Luca		U.Laval	luca.freschi@bio.ulaval.ca
Gagnon Yves	T	UdeM	y.gagnon@umontreal.ca
Girard Nicolas		UdeM	nicolas1.1girard@gmail.com
Grenier Jean-Christophe	T	UdeM	jean-christophe.grenier@umontreal.ca
Hallett Michael	Keynote	McGill	hallett@mcb.mcgill.ca
Hussin Julie	T	UdeM	Julie.hussin@umontreal.ca
Ishii Tetsu	P	UdeM	tetsu.ishii@umontreal.ca
Karp Peter	Keynote	SRI	pkarp@ai.sri.com
Kiethega Georgette	Volunteer	UdeM	georgette.kiethega@umontreal.ca
Kleinman Claudia Laura	Jury	McGill	cl.kleinman@mcgill.ca
Lampron Philipe	Volunteer	UdeM	philipe.lampron@umontreal.ca
Lalonde Emilie	P	McGill	emilie.lalonde3@mail.mcgill.ca
Langevin Sébastien	bin6000	UdeM	sebastien.langevin@umontreal.ca
Lang B. Franz	Chair, Jury	UdeM	Franz.lang@umontreal.ca
Lartillot Nicolas	Chair, Jury	UdeM	nicolas.lartillot@umontreal.ca
Lalonde Emilie CANCELLED	P	McGill	emilie.lalonde3@mail.mcgill.ca
Lavallée-Adam Mathieu	T	McGill	mathieu.lavallee-adam@mail.mcgill.ca
Layeghifard Mehdi		UQAM	layeghifard.mehdi@courrier.uqam.ca
Leclercq Mickael		UQAM	leclercq.mickael@courrier.uqam.ca
Leducq Jean-Baptiste	P	U.Laval	leducq.jean-baptiste@laposte.net
Lord Étienne	P	UQAM	lord.etienne@courrier.uqam.ca
Lussier Anne-Qing		UdeM	dontfall_bewall@sympatico.ca
Makarenkov Vladimir	Jury	UQAM	makarenkov.vladimir@uqam.ca
Malekpour Seyed Amir	P	UdeM	seyed@bch.umontreal.ca
Marois François		U.Laval	francois-christophe.marois-blanchet.1@ulaval.ca
Men Dararoth	bin3002	UdeM	dararoth.men@umontreal.ca
Meunier Elaine	Volunteer	UdeM	elaine.meunier@umontreal.ca
Michaut Magali	P,T	U.Toronto	magali.michaut@utoronto.ca

Minoiu Ioana	Volunteer	UdeM	ioana.minoiu@umontreal.ca
Moreira Sandrine	P, Organiz.	UdeM	sandrine.moreira@umontreal.ca
Morozov Boris	bin3002	UdeM	boris.coldman@gmail.com
Mousseau Normand	Keynote	UdeM	normand.mousseau@umontreal.ca
Murua Alejandro	Jury	UdeM	murua@DMS.umontreal.ca
Nga’mbe Clarisse			clarissen@hotmail.com
Nikbakht Hamid	p	Concordia	h.nikbakht@gmail.com
Noel Audrey	Volunteer	UdeM	audrey.noel@umontreal.ca
O'Brien Emmet	Jury	Concordia	emmetaobrien@gmail.com
O'Reilly Patrick	P	UdeM	patrick.oreilly@umontreal.ca
Pacis Alain		UdeM	vapacis@gmail.com
Parto Sahar	P	UdeM	sahar.parto@umontreal.ca
Philippe Hervé	Jury	UdeM	herve.philippe@umontreal.ca
Poujol Raphael	T	UdeM	raphael.poujol@umontreal.ca
Ragonnet-Cronin Manon	P	U.Ottawa	mrago029@uottawa.ca
Roure Béatrice	T	UdeM	beatrice.roure@umontreal.ca
Rousseau Mathieu	T	McGill	mathieu.rousseau3@mail.mcgill.ca
Samba Maty Laye	bin3002	UdeM	matysa28@hotmail.fr
Serpa-Ibanez Roman Felipe	bin6000	UdeM	roman.serpa@umontreal.ca
Shateri-Najafabadi Hamed	P	U. McGill	hamed.shaterinajafabadi@mail.mcgill.ca
Song Carl	T	U.Toronto	carl.song@utoronto.ca
St-Onge Karine	P	UdeM	chatonn@hotmail.com
St-Pierre Jean-Francois	P	UdeM	jf.st-pierre@umontreal.ca
Szathmary Laszlo		UQAM	jabba.laci@gmail.com
Tahiri Nadia		UQAM	tahiri.nadia@courrier.uqam.ca
Tapsoba Franck	bin3002	UdeM	franck.tapsoba@umontreal.ca
Théroux Jean-François	P	UdeM	jean-francois.theroux@umontreal.ca
Tillier Elisabeth	Sci.com	U.Toronto	e.tillier@utoronto.ca
Torres-Quiroz José-Francisco		U.Laval	jose-francisco.torres-quiroz.1@ulaval.ca
Tremblay-Savard Olivier	T	UdeM	olivier.tremblay-savard@umontreal.ca
Valencia Alfonso	Keynote	CNIO	valencia@cnio.es
van Weringh Anna	T	U.Ottawa	avanw070@uottawa.ca
Vello Emilio	P	UdeM	emilio.damian.vello@umontreal.ca
Yan Yifei	Volunteer	UdeM	yifei.yan@umontreal.ca

VIPs

Baron Christian	Head of Biochemistry, UdeM
Hubert Joseph	Vice Rector, Research, UdeM
Jonas-Cedergren Henrietta	Emeritus Professor, UQAM
Turcotte Patrice	Head of Computer Science (DIRO), UdeM