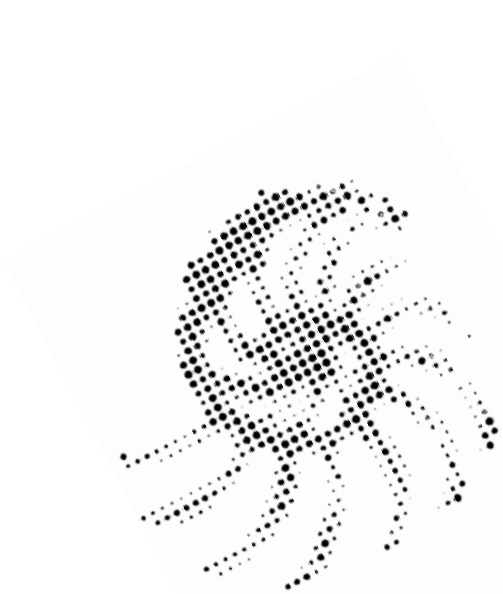
COLLOQUE BIO-INFORMATIQUE

ROBERT-CEDERGREN



Programme

Université de Montréal
14-15 novembre 2006

Présentations orales et Affiches
Poster and oral presentations

http://www.centrerc.umontreal.ca/colloqueannuel.html

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

Université de Montréal

Bienvenue au 3e colloque bio-informatique Robert-Cedergren !

Ce colloque se veut le rendez-vous annuel de la communauté universitaire oeuvrant en bio-informatique. L'objectif principal est de partager les derniers développements en ce domaine par le biais d'un concours d'affiches et de présentations orales et de rendre compte de l'importance grandissante de la bio-informatique dans les sciences de la vie.

En cette troisième édition du colloque, les conférenciers invités sont Stephen Altschul, Ph.D., Computational Biology Branch, National Center for Biotechnology Information, National Institutes for Health (USA), bien connu pour avoir développé le logiciel Basic Local Alignment Search Tool (BLAST) ainsi que Brian Golding, Ph.D., professeur au Département de Biologie et directeur du laboratoire de bio-informatique à l'Université McMaster, Hamilton, Ontario.

Quatorze présentations orales et vingt affiches seront en lice dans quatre catégories. L'Université ayant le meilleur classement remportera le prix Robert-Cedergren.

Cette sculpture d'acrylique représente l'outil de travail par excellence du bio-informaticien : l'ordinateur. Sur l'écran, une molécule d'ARN de transfert, thème de recherche cher à Robert Cedergren.

Les prix individuels seront décernés dans les catégories suivantes :

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

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

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

Welcome to the 3rd annual Robert Cedergren Bioinformatics Colloquium!

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

Keynote speakers will be Stephen Altschul, PhD, NCBI, NLM, NIH, Computational Biology Branch, well known for his contribution in the Basic local alignment search tool (BLAST) program, and Brian Golding, Ph.D. Professor and Director of Computational Biology Laboratory in McMaster University, Hamilton, Ontario.

This year, 14 oral presentations and 20 posters will compete in 4 categories. The university with the best ranking will win the Robert Cedergren Award.

This sculptur made of acrylic represents the bioinformatician's tool : the computer. On the screen is displayed, a tRNA molecule, one of the preferred research themes of Robert Cedergren.

Individual awards will be given in the following categories:

	Best oral presentations	Best posters
MSc	1000 $	500 $
PhD	1000 $	500 $

Enjoy this 3rd annual Robert-Cedergren Bioinformatics Colloquium!

Gertraud Burger, Ph.D.
Leader
Bioinformatics graduate programs

Comités/Committees

Comité d'organisation / Organizing Committee

Gertraud Burger
Marie Robichaud

Arbitres / Referees

Stéphane Aris-Brossou (U. Ottawa)
Sylvie Hamel (U. Montréal)
Sébastien Lemieux (U. Montréal)
Sabin Lessard (U.Montréal)
Hervé Philippe (U. Montréal)
Serguei Chteinberg (U. Montréal)
Marcel Turcotte (U. Ottawa)

Responsables de séance / Session Chairs

François-Joseph Lapointe
François Major
Normand Mousseau
Daniel Sinnett

Renseignements généraux / General information

Accueil / Registration

L'accueil des participants se fera au Hall d'honneur du pavillon Roger-Gaudry (anciennement Pavillon principal), le mardi 14 novembre dès 8 h 30. Les insignes d'identification vous seront remis à ce moment.

The registration office is located in the Honor Hall in the Roger-Gaudry Building (previously Main Building). Your identification badge will be available from 8:30 am, November 14.

Pauses santé et cocktail / Coffee breaks and cocktail

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

Coffee breaks, lunches and the cocktail party will be served in the Honor Hall.

Horaire du mardi 14 novembre 2006

9 h 15 **Ouverture du colloque : Jacques Turgeon, vice-recteur à la recherche Université de Montréal**

9 h 30 **Conférence 1 : Brian Golding, McMaster University**
«Laterally transferred genes must quickly become useful to the host before they are lost: An examination of rates and patterns of lateral transfer in bacteria»

10 h 30 **OM1 : Véronique Lisi (UdeM)**
The triloop motif structure

AM1 : Martin Smith (UdeM)
Trypanosomatid transcriptomics : Predicting poly-A sites and 5' splice juntions of polycistronic mRNA

AM2 : Pascal St-Onge (UdeM)
Détection et caractérisation des interactions gène-gène dans la susceptibilité à la leucémie de l'enfant: approche statistique et informatique

AM3 : Malika Aïd (UdeM)

11 h 00 Pause + Posters session

11 h 30 **OM2 : François Belleau (Laval)**
How to RDFize bioinformatics databases : the Bio2RDF's recipe

AM4: Sébastien Christin (UdeM)
Recherche de snoRNA de type boîte C/D en corrélation avec le signal de reconnaissance de l'enzyme Rnt1p

AM5: Mohamed Tikah Marrakchi (UdeM)
Helix Explorer: A new database of protein structures

AM6 : Pascal Bachand (UdeM)
Prédiction de la susceptibilité à l'hypertension dans une population canadienne française - analyse d'association des haplotypes par apprentissage machine

OM = Oral M.Sc.
OD = Oral Ph.D.

AM = Affiche M.Sc. Poster
AD = Affiche Ph.D. Poster
APD = Affiche Postdoc Poster

Oral presentations : M-415

Posters : Hall d'honneur

12 h 00 **OM3 : Hamsa Tadepally (UdeM)**
Evolution of C2H2-zinc finger genes in mammalian genomes

AM7 : Jean-François St-Pierre (UdeM)
Exploration des surfaces d'énergies avec la méthode ART nouveau et FRODA

AD13 : Tetsu Ishii (UdeM)
Selenocysteine tRNA (tRNAsec) delivers a rare amino acid selenocysteine in response to a UGA stop codon and a certain downstream stem loop signal in mRNA

AD14: François Boulay (UdeM)
Modeling tRNA conformity at the atomic level

12 h 30 Lunch + Posters session

13 h 30 **OD1: Zaky Adam (Ottawa)**
Inférence des Arbres Phylogénomique en Utilisant des Operations DCJ

14 h 00 **OD2 : Abdoulaye Baniré Diallo (McGill) (annulé)**
Inférence du scenario d'indels le plus vraisemblable

14 h 30 Pause + Posters session

16 h 00 **OD3 : Jean-Eudes Duchesne (UdeM)**
A seeding method for RNA local alignment search tools

16 h 30 **OD4 : Mathieu Lajoie (UdeM)**
Evolution of Tandemly Repeated Genes through Duplication and Inversion

17 h 00 Rafraichissements + Posters session

OM = Oral M.Sc.
OD = Oral Ph.D.

AM = Affiche M.Sc. Poster
AD = Affiche Ph.D. Poster
APD = Affiche Postdoc Poster

Oral presentations : M-415

Posters : Hall d'honneur

Horaire du mercredi 15 novembre 2006

9 h 00 **Conférence 2 : Stephen Altschul, NCBI**
«Retrieval Accuracy, Statistical Significance, and Compositional Similarity in Protein Sequence Database Searches»

10 h 00 **OD5: Véronique Campbell (UdeM)**
Assessing congruence among distance matrices prior to phylogenetic analysis

AD1 : Claudia Laura Kleinman (UdeM)
Statistical potentials for phylogeny and protein design

AD2 : Nicolas Rodrigue (UdeM)
Markov chain Monte Carlo algorithms for likelihood and Bayesian phylogenetic

AD3 : Béatrice Roure (UdeM)
SCaFoS: Selection, Concatenation and Fusion of Sequences for phylogenomics

10 h 30 **OD6 : Sivakumar Kannan (UdeM)**
Evaluation ORF Function Predictions using Domain-specific knowledge

AD4 : Richard Daigle (Laval)
Characterization of the structure and dynamics of the truncated hemoglobin trHbN

AD5 : Olivier Fisette (Laval)
Backbone dynamics of β-lactamase TEM-1 at the crossroads of molecular dynamics and NMR spectroscopy

AD6 : Jean-Eudes Duchesne (UdeM)
A seeding method for RNA local alignment search tools

11 h 00 Pause-santé

11 h30 **OD7: Wei Xu (Ottawa)**
The Distance Between Randomly Constructed Genomes

AD7: Yaoqing Shen (UdeM)
Improving the prediction of mitochondrial proteins via the integration of available tools for subcellular localization

OM = Oral M.Sc.
OD = Oral Ph.D.

AM = Affiche M.Sc. Poster
AD = Affiche Ph.D. Poster
APD = Affiche Postdoc Poster

Oral presentations : M-415

Posters : Hall d'honneur

AD8 : Daniel Darmon (UdeM)
The evolutionary transition from protists to Metazoa: mitochondrial genome organization and phylogenomic analyses based on nuclear and mitochondrial genes

AD9: Yan Zhou (UdeM)
Evaluation of the models handling heterotachy in phylogenetics inference

12 h 00 **OD8 : Alizerah Shaneh (McGill)**
Structural Basis of ATP Binding Affinity to RNA Editing Ligase 1

AD10 : Jean-Philippe Doyon (UdeM)
Inferring a duplication and speciation history from a gene tree

AD11 : Yu Liu (UdeM)
A likelihood ratio test to identify fast-evolving sites in protein sequence, and its application on phylogenomic analysis

AD12 : Jian Hong Chen (UdeM)
Sliding Between RNA Helical Elements Revealed by a Molecular Dynamics Simulation Study

12 h 30 Dîner + Posters session

13 h 30 **OD9 : Lilianne Dupuis (UdeM)**
Étude de la flexibilité des protéines par simulations informatiques multi-échelles

14 h 00 **OD10 : Claudia Kleinman (UdeM)**
Statistical potentials for phylogeny and protein design

14 h 30 **OD11 : Nicolas Rodrigue (UdM)**
Bayesian modeling of protein coding sequence evolution

15 h 00 Pause + Posters Session

16 h 30 Remise des prix : Gertraud Burger
16 h 45 Mot de clôture : Jacques Turgeon, Vice-recteur à la recherche

17 h 00 Cocktail et bouchées

OM = Oral M.Sc.
OD = Oral Ph.D.

AM = Affiche M.Sc. Poster
AD = Affiche Ph.D. Poster
APD = Affiche Postdoc Poster

Oral presentations : M-415

Posters : Hall d'honneur

CONFÉRENCES

Brian Golding, Ph.D.
Department of Biology
McMaster University

Laterally transferred genes must quickly become useful to the host before they are lost: An examination of rates and patterns of lateral transfer in bacteria.

Weilong Hao, Pradeep Marri and G. Brian Golding

A substantial number of genes can be laterally transferred within a very short period of evolutionary time. Detecting or inferring gene insertions/deletions is of interest because such information provides insights into bacterial genome evolution and speciation. We have used a simple maximum likelihood method to infer these lateral transfer events according to the phylogenetic history of the taxa. I will discuss studies that we have done to examine the rates and patterns of laterally transferred genes to/from the genomes genomes of closely related bacteria. The presence or absence of genes from each genome were cataloged. It is shown that whole gene insertions/deletions in genomes occur at rates comparable to or greater than the rate of nucleotide substitution and that higher insertion/deletion rates are often inferred to be present at the tips of the phylogeny with lower rates on more ancient interior branches. Recently transferred genes are under faster and relaxed evolution compared with more ancient genes. Many of the lineage-specific insertions might be lost quickly during evolution and perhaps some of the genes inserted by lateral transfer are niche specific.

Stephen Altschul, Ph.D.
NCBI

Retrieval Accuracy, Statistical Significance, and Compositional Similarity in Protein Sequence Database Searches

Protein sequence database search programs may be evaluated both for their retrieval accuracy and for the accuracy of their statistical assessments of reported alignments. However, methods for improving statistical accuracy can degrade retrieval accuracy by discarding compositional evidence of sequence relatedness. This evidence may be preserved by combining essentially independent measures of alignment and and compositional similarity into a unified measure of sequence similarity. A version of the BLAST program, modified to employ this new measure, outperforms the baseline program in both retrieval and statistical accuracy on ASTRAL, a SCOP-based test set.

PRÉSENTATIONS ORALES / ORAL PRESENTATIONS

OM1 : Véronique Lisi (UdeM)

The triloop motif structure

A systematic and exhaustive search of all RNA triloops present in the X-ray structures (res. < 3Å) was made. The structural classification of these triloops is a rooted tree containing 8 topological classes, 26 subclasses, 54 structures, 102 specimens, and 917 triloop instances. It produced a useful sequence to structure relation reference for RNA 3 D structure modeling and sequence-based experimental approaches. Specimens were found in most RNAs and in many different contexts and many specimens are found within more complex motifs that include inserted nucleotides.

OM2 : François Belleau (UdeM) (présentation en français)

How to RDFize bioinformatics databases : the Bio2RDF's recipe

The integration of bioinformatics knowledge can be achieved through the inherent aggregate capabilities of the Semantic web technologies based on the use of RDF. Bio2RDF project (http://bio2rdf.org), a new knowledge atlas for human and mouse, has also the same goal.

During our presentation we will how we converted NCBI's OMIM, GeneID, GENBANK and PubMed documents from XML to RDF. We will also show how to convert chEBI MySQL database and UniProt text format to RDF.

The source code of the JSP RDFizer with a functional Sesame server will be made available at the following web site http://sourceforge.net/projects/bio2rdf.

OM3 : Hamsa Tadepally (UdeM)

Evolution of C2H2-zinc finger genes in mammalian genomes

Hamsa Tadepally, Gertraud Burger and Muriel Aubry.

Zinc finger genes of the C2H2 type (C2H2-ZNF) represent the largest gene family of transcription factors in the human genome. Using an extensive homology search we identified 711 human genes often grouped into clusters. Analysis of homologous C2H2-ZNF clusters, in chimpanzee, dog and mouse genomes in regions syntenic to human, indicate differential duplication and loss of genes and finger motifs in addition to shuffling of domains leading to different C2H2-ZNF repertoires in mammalian genomes.

OD1 : Zaky Adam (Ottawa)

Inférence des Arbres Phylogénomique en Utilisant des Operations DCJ

The universal double cut and join (dcj) operation accounts for inversions, translocations, fissions, fusions and generalized transpositions. This feature justifies the fact that dcj is an efficient tool in measuring the minimum distance between two uni or multichromosome genomes. We first use dcj to reconstruct gene orders of the median ancestor of three genomes, and then we use the median ancestor to reconstruct gene orders of the ancestral genomes in a given tree. The number of these operations needed to convert one genome to another is used as a distance measure.

OD2 : Abdoulaye Baniré Diallo (McGill)

Inférence du scenario d'indels le plus vraisemblable

La reconstruction des scénarios d'insertions et de déletions (indels) à partir d'un alignement de séquences constitue une des étapes cruciales de la reconstruction de sequences ancestrales. Bien que ce problème ait déjà été abordé d'une façon parcimonieuse, il est préférable de trouver la reconstruction la plus vraisemblable capapable d'expliquer les brèches observées dans l'alignement. Nous proposons une solution combinant un Modèle de Markov de Caché (MMC) et un arbre phylogénétique. La méthode est illustrée sur des séquences de mammifères provenant du projet ENCODE.

OD3 : Jean-Eudes Duchesne (UdeM)

A seeding method for RNA local alignment search tools

During the past few decades the RNA world has been shaken back alive with the discovery of their catalytic capabilities. It is now apparent that RNA molecules are not solely relegated to the menial task of translation into proteins but are also a player in the fast track world of gene regulation. Unfortunately, tools have not developed as fast as the renewed interest in the molecule. Some techniques are capable of searching very detailed descriptions of RNA families while others try to discover new structural features to refine structural search. But while the data on RNA is growing exponentially, few efforts are made to speed the search process for large scale deployment. Methods like Blast and Patternhunter are highly successful for protein search but currently have no RNA counterpart. Traditionally, these methods have done relatively poorly when dealing with RNA molecules. Our work builds on the concept of hashing methods for speeding up searches in large scale databases by building seeding elements doesn't discriminate against RNA's special features. It was shown that including "don't care" characters into the seed can either speed up hashing methods for biological sequences or higher the overall sensibility of the method. We further generalize this approach by creating a variable sized seed that can include gaps of undefined elements. With this method, we can significantly speed up large scale RNA search while maintaining an acceptable sensitivity ratio.

OD4 : Mathieu Lajoie

Evolution of Tandemly Repeated Genes through Duplication and Inversion

Plusieurs études récentes considèrent le problème de la reconstruction d'une histoire évolutive pour une famille de gènes répétés en tandem. Cependant, les méthodes proposées sont souvent inapplicables car le modèle actuel de duplication ne peut expliquer les différentes orientations transcriptionnelles que l'on retrouve au sein de plusieurs familles. Nous proposons le premier modèle de duplication permettant les inversions ainsi qu'une application sur un jeu de données biologiques.

OD5: Véronique Campbell

Assessing congruence among distance matrices prior to phylogenetic analysis

In phylogenetic analysis, different genes or different types of characters are often combined in a single analysis to infer evolutionary relationships among taxa. CADM (Congruence Among Distance Matrices) can be used to test whether these different data matrices contain congruent information. In this study, we present simulations to estimate the type 1 error rate and power of CADM. Contrarily to the other tests of incongruence, CADM allows comparison of more than two matrices at a time and can be applied to distance matrices.

OD6 : Sivakumar Kannan (UdeM)

Evaluation ORF Function Predictions using Domain-specific knowledge

Evaluating the machine learning based function predictions of hypothetical proteins is a challenge task. Therefore, we have developed evaluation criteria based on domain-specific knowledge. Using these criteria, we were able to rank predictions and thus identifying the most likely working hypotheses for experimental validation.

OD7 : Wei Xu (Ottawa)

The Distance Between Randomly Constructed Genomes

To see whether the comparison of two genomes contains some signal of the evolutionary processes responsible for their divergence, aside from the total number of breakpoints between their synteny blocks, we should compare the rearrangement analysis of these genomes with the same analysis on pairs of randomized genomes with the same number of breakpoints. In previous papers, we have worked out the statistical properties of random genomes consisting of one or more circular chromosomes [1], and those of two random genomes containing the same number c of linear chromosomes [2]. The later paper concentrated on showing that the number of circular chromosomes inevitably associated with random linear chromosomes is very small with realistic numbers of chromosomes. It only included a rough estimation of the statistical properties of the linear chromosomes. This presentation introduces a new way of representing the comparison of linear genomes, instead of the numerous "chromosomal caps" used in other treatments. This facilitates a more rigorous treatment of the case of linear chromosomes, including the more realistic situation where the number of linear chromosomes may be different \chi_1 and \chi_2) in the two genomes being compared.

OD8 : Alireza Shaneh (McGill)

Structural Basis of ATP Binding Affinity to RNA Editing Ligase 1

RNA Editing is a post-transcriptional modification of mitochondrial mRNAs in trypanosomatid pathogens. The process generates mature funtional mRNAs and is catalyzed by a large multi-protein complex machinery called editosome. It has been shown that RNA editing is essential for survival of the disease causing bloodstream stage of Trypanosoma brucei, thereby making it a rational target for possible inhibitors. RNA editing ligase 1 (REL1) is a key enzyme component of the editosome complex. In this study, the crystallography information of the catalytic domain of T. brucei REL1 (TbREL1) in complex with ATP at 1.2 Angstrom was used as the main source for the molecular dynamics simulation. In this presentation, the ATP affinity to the binding site of TbREL1 will be discussed. The preliminary results of an energy minimization cascade through the cleaned structure of TbREL1 in the presence or absence of ATP will also be demonstrated. Such a study effectively investigates the local minimum level of potential energy surface for the residues with certain threshold radius (6 Angstrom in this research) around the binding site of TbREL1 and provides insight into the future design of the chemical compounds which can potentially target the essential RNA editing process across major trypanosomatid pathogens.

OD9 : Lilianne Dupuis (UdeM)

Étude de la flexibilité des proteins par simulations informatiques multi-échelles

Étude de la flexibilité des protéines par simulations informatiques multi-échelles. L'étude de la flexibilité des protéines nous aide à comprendre leur fonction. Une protéine existe en plusieurs formes selon son état de liaison. Des méthodes numériques tentent de déduire les trajectoires entre celles-ci, mais les temps de calcul sont trop élevés. Nous développons donc des représentations moléculaires à différentes échelles de réalité, exploitant le fait que de grandes zones de la protéine se meuvent en un seul bloc. Le temps de calcul se consacre alors surtout aux zones flexibles.

OD10 : Claudia Laura Kleinman

Statistical potentials for phylogeny and protein design

Proteins contain in their sequence all the information to fold into the native structure, but understanding the way this information is coded remains elusive. This is, in part, because natural sequences are the result of a complex evolutionary process involving multiple factors. In order to characterize the constraints on protein sequences exclusively due to the structure, we present a new set of statistical potentials, i.e. energy functions where the parameters are learnt from a database of known protein structures.

OD11 : Nicolas Rodrigue (UdM)

Bayesian modeling of protein coding sequence evolution

Mechanistic statistical models are attempts at formulating holistic descriptions that explain causative phenomena producing a set of observations. In 1994, Muse & Gaut (MG) and Goldman & Yang (GY) proposed such strategies for modeling the evolution of protein coding genes in a phylogenetic context; these models recognize the genetic code, defining a state space consisting of the 61 sense codons. Numerous variations and extensions have since been developed, yielding powerful tools for characterizing selective pressures across codon sites and across lineages. However, several long-standing issues regarding subtle differences in the specification of MG-type and GY-type codon models have remained largely unaddressed, for the most part due to the theoretical difficulties of comparing non-nested statistical models in the maximum likelihood paradigm. In this work, we report the first Bayesian measurements of codon model fit, comparing MG and GY formulations using Bayes factors. We also propose a new model, building on the MG formulation, which accounts for codon usage bias, while more closely subscribing to the mechanistic modeling philosophy. Using two real datasets, our findings indicate that MG-type models are preferred over GY models, and that our new model provides the best fit of all. Altogether, these results call for a larger scale study of alternative codon models from a Bayesian standpoint, and suggest future research directions aimed toward building a hierarchical description of molecular evolution, which acknowledges the underlying nucleotidic substitution process, inherent biases of the translational machinery, and the overall phenotypic effects of the encoded proteins.

AFFICHES / POSTERS

AM1 : Martin Smith (UdeM)

Trypanosomatid transcriptomics : Predicting poly-A sites and 5' splice juntions of polycistronic mRNA

Leishmania and other trypanosomatids have tandemly arranged genes and little (if any) transcriptional regulation. We wish to predict the 5' and 3' RNA transcript boundaries for automated large-scale expression analysis. Using published EST data, we developed scoring schemes that perdict mRNA boundaries from annotated genomic sequences.

AM2 : Pascal St-Onge (UdeM)

Détection et caractérisation des interactions gène-gène dans la susceptibilité à la leucémie de l'enfant: approche statistique et informatique

La leucémie lymphoblastique aiguë (LLA) est le cancer le plus fréquent chez les enfants, comptant pour 25-35% des cas de cancers pédiatriques dans le monde occidental. La LLA est considéré comme un modèle de maladie complexe où les effets de séries de gènes de faibles pénétrances sont modulés par des facteurs externes. L'étude de ces interactions représente un défi de taille en raison de l'importante dimensionnalité des échantillons de données. Pour attaquer ce problème, nous proposons une approche basée sur les réseaux de neurones qui sera comparée à la méthode de référence Multifactor Dimensionality Reduction (MDR).

AM3 : Malika Aïd

La régulation de l'expression des gènes chez les eucaryotes fait intervenir des activateurs ou répresseurs transcriptionnels se liant à des séquences d'ADN spécifiques. Cette technique, combinée au séquençage systématique des sites d'ADN fixés, présente aussi le potentiel de révéler de nouveaux modes de liaison de ces facteurs. En effet, la liaison indirecte de facteurs de transcription à l'ADN par l'intermédiaire d'autres facteurs de transcription (« tethering ») est de plus en plus documentée. Cependant, l'importance de ces mécanismes d'action à l'échelle du génome reste à déterminer. L'identification des sites de liaison à l'ADN permet une meilleure compréhension du mécanisme de régulation via les facteurs de transcription.Notre objectif présent est de développer un système intégré d'identification de sites dans des jeux de données de ChIP permettant de minimiser le taux de faux positifs pour faciliter la validation expérimentale subséquente, qui représente une étape limitante.

AM4 : Sébastien Christin

Recherche de snoRNA de type boîte C/D en corrélation avec le signal de reconnaissance de l'enzyme Rnt1p

Sébastien Christin, Nadia El-Mabrouk et Sherif Abou Elela

Les snoRNAs de type boîte C/D sont des séquences d'une taille variant de 70 à 250 nucléotides. Ces molécules sont impliquées dans la méthylation de certains nucléotides sur les ARNs ribosomaux et les snRNAs. Cette classe d'ARN non codant possède quelques caractéristiques universelles: boîte C (consensus AUGAUGA), boîte D (consensus CUGA), et l'hélice terminale d'environ 4 à 8pb fermant la molécule.

La recherche de snoRNAs à l'aide de ces caractéristiques n'est pas adéquate. En effet, sur un génome de la taille de *S.cerevisiae,* on obtient des milliers de possibilités. Par contre, un motif de clivage se trouve dans le voisinage de la séquence codante de snoRNA. Étant reconnu par l'enzyme RNT1P, ce motif intervient dans la maturation du snoRNA. Une caractérisation et une fouille de ce signal de reconnaissance dans le génome permettrait d'avoir un critère de sélection supplémentaire pour la découverte de snoRNAs.

Plusieurs méthodes ont été utilisées pour caractériser davantage ces motifs clivés par l'enzyme RNT1P: statistique d'information mutuelle, recherche de motif simple, analyse de la structure d'hélice, apprentissage machine (SVM), analyse du niveau d'expression génétique.

La méthode, à ce jour, permet de trouver 12 nouveaux candidats snoRNAs, tout en récupérant 14 des 21 vrais snoRNAs de type boîte C/D répertoriés dans la littérature. La recherche sur le motif de clivage effectuée, tout en étant très restrictive du fait qu'on veut obtenir peu de candidats snoRNAs, permet tout de même de retrouver 65% (32/49) de tous les motifs répertoriés.

AM5 : Mohamed Tikah Marrakchi

Helix Explorer: A new database of protein structures

Despite being only at its infancy stage, protein design and engineering is a discipline that would certainly benefit from a better access to the information contained in the ever growing Protein Data Bank (PDB) structures repository. We introduce hereby Helix Explorer, a searchable database of protein secondary structure elements that provides protein designers with new means of querying the PDB and, through its dynamic Web interface, makes protein structures analysis more accessible.

AM6 : Pascal Bachand

Prédiction de la susceptibilité à l'hypertension dans une population canadienne française - analyse d'association des haplotypes par apprentissage machine

La haute pression artérielle (hypertension) est une condition très commune qui affecte près du quart de la population adulte en Amérique du nord. Cette dernière peut emmener plusieurs complications affectant à divers degré la santé du patient mais malheureusement, un nombre important des gens touchés ne sont pas conscient de leur condition. Notre projet vise à détecter la susceptibilité à l'hypertension à partir du profile haplotypique d'un individu, en réalisant préalablement une analyse d'association sur les données pangénomiques de 455 patients originaires du Saguenay - Lac St-Jean. Cette étude est la poursuite d'un projet débuté en 1999 portant sur les effets fondateurs génomiques de cette population canadienne française. Le défi du projet consiste à réaliser cette association malgré la dimension importante des données (58,000 SNP) par rapport au nombre d'échantillons (455 patients: 237 hypertendus, 218 contrôles). L'approche informatique employée consiste d'abord à sélectionner un nombre réduit de blocs haplotypiques en fonction de divers algorithmes de classement, pour ensuite réaliser une classification par apprentissage machine sur le phénotype d'hypertension. Plusieurs algorithmes ont été testés, comparés et ultimement combinés au cours de ce projet, et l'affiche présentée résume les méthodes, résultats et conclusions notés jusqu'à présent.

AM7 : Jean-François St-Pierre (UdeM)

Exploration des surfaces d'énergies avec la méthode ART nouveau et FRODA

Au cours des cinq dernières années, la méthode d'activation-relaxation (ART nouveau) a été utilisée avec succès dans l'étude des trajectoires de repliement de petites protéines et dans l'agrégation de peptides. À fin d'optimiser l'efficacité de cette méthode d'exploration des surfaces d'énergie, nous avons implanté l'algorithme des corps fantômes FRODA pour diminuer l'impacte des degrés de liberté de haute fréquence. Cette affiche vous présent les différents degrés d'intégration de FRODA dans les composantes d'ART nouveau.

AD1 : Claudia Laura Kleinman

Statistical potentials for phylogeny and protein design

Proteins contain in their sequence all the information to fold into the native structure, but understanding the way this information is coded remains elusive. This is, in part, because natural sequences are the result of a complex evolutionary process involving multiple factors. In order to characterize the constraints on protein sequences exclusively due to the structure, we present a new set of statistical potentials, i.e. energy functions where the parameters are learnt from a database of known protein structures.

AD2 : Nicolas Rodrigue

Markov chain Monte Carlo algorithms for likelihood and Bayesian phylogenetic analysis

In recent years, the advent of Markov chain Monte Carlo (MCMC) techniques, coupled with modern computational capabilities, has enabled the study of evolutionary models without a closed form solution of the likelihood function, offering greater flexibility in the exploration of different modeling strategies. However, current Bayesian MCMC applications can incur significant computational costs, as they are based on a full sampling from the posterior probability distribution of the parameters of interest. Here, we draw attention as to how MCMC techniques can be embedded within normal approximation strategies for more economical statistical computation. The overall strategy is based on an estimate of the first and second moments of the likelihood function, as well as a maximum likelihood estimate. Through examples, we review several MCMC-based methods used in the statistical literature for this purpose, applying the approaches to constructing posterior distributions under non-analytical evolutionary models relaxing the assumptions of rate homogeneity, and of independence between sites. Finally, we use the procedures for conducting Bayesian model selection, based on Laplace approximations of Bayes factors, which we find to be accurate and computationally advantageous. Altogether, the methods we expound here, as well as other related approaches from the statistical literature, should prove useful when investigating increasingly complex descriptions of molecular evolution, alleviating some of the difficulties associated with non-analytical models.

AD3 : Béatrice Roure

SCaFoS: Selection, Concatenation and Fusion of Sequences for phylogenomics

Analyzing a large number of genes and species is becoming a standard approach to resolve difficult phylogenetic questions. However, the creation of large datasets implies dealing with a lot of problems (multiple copies of a gene, missing genes, partial sequences, paralogous genes) SCaFoS quickly assembles phylogenomic datasets with maximal phylogenetic information, allowing selection of species and genes while adjusting the amount of missing data. Through a judicious selection of data, SCaFoS may reduce potential artefacts in phylogenetic inferences, especially for species with a high evolutionary rate.

AD4 : Richard Daigle (Laval)

Characterization of the structure and dynamics of the truncated hemoglobin trHbN

The truncated hemoglobin trHbN protects the aerobic respiration of Mycobacterium tuberculosis by oxidizing nitric oxide into nitrate ion. The main objective of our research is to characterize the structure and the dynamic properties of trHbN. To achieve this goal we performed molecular dynamic simulations under various conditions. We present results obtained from simulations performed on the oxygenated and deoxygenated trHbN forms in which the conformational space of trHbN was explored.

AD5 : Olivier Fisette (Laval)

Backbone dynamics of β-lactamase TEM-1 at the crossroads of molecular dynamics and NMR spectroscopy

Backbone dynamics of β-lactamase TEM-1 (order parameters and correlation times of amide N-H vectors) are caracterised using molecular dynamics simulations. Results are compared to those obtained using NMR spectroscopy. We are thus able to confirm and refine the abstract models used for the analysis of NMR spectroscopy data, and to provide a physical interpretation of these models. The high rigidity of TEM-1's backbone is related to its exceptional enzymatic efficiency and adaptability.

AD6 : Jean-Eudes Duchesne (UdeM)

A seeding method for RNA local alignment search tools

During the past few decades the RNA world has been shaken back alive with the discovery of their catalytic capabilities. It is now apparent that RNA molecules are not solely relegated to the menial task of translation into proteins but are also a player in the fast track world of gene regulation. Unfortunately, tools have not developed as fast as the renewed interest in the molecule. Some techniques are capable of searching very detailed descriptions of RNA families while others try to discover new structural features to refine structural search. But while the data on RNA is growing exponentially, few efforts are made to speed the search process for large scale deployment. Methods like Blast and Patternhunter are highly successful for protein search but currently have no RNA counterpart. Traditionally, these methods have done relatively poorly when dealing with RNA molecules. Our work builds on the concept of hashing methods for speeding up searches in large scale databases by building seeding elements doesn't discriminate against RNA's special features. It was shown that including "don't care" characters into the seed can either speed up hashing methods for biological sequences or higher the overall sensibility of the method. We further generalize this approach by creating a variable sized seed that can include gaps of undefined elements. With this method, we can significantly speed up large scale RNA search while maintaining an acceptable sensitivity ratio.

AD7 : Yaoqing Shen (UdeM)

Improving the prediction of mitochondrial proteins via the integration of available tools for subcellular localization

More than a dozen computational tools are available to predict proteins localized in mitochondria. But their predictions for a given sequence often disagree, which makes it difficult to choose the correct answer. In order to make highly reliable predictions, we developed a new classifier based on decision tree, which integrated 12 available tools for the prediction of subcellular localization and transmembrane domains. The new classifier has a sensitivity of 86% and precision of 96%, which are significantly better than any of the individual tools.

AD8 : Daniel Darmon (UdeM)

The evolutionary transition from protists to Metazoa: mitochondrial genome organization and phylogenomic analyses based on nuclear and mitochondrial genes

D. Darmon[1], Y. Liu[1], E. Steenkamp[2], G. Burger[1] and B. F. Lang[1]

[1]Robert Cedergren Centre for Bioinformatics and Genomics, Canadian Institute for Advanced Research Département de biochimie, Université de Montréal, Montréal, Québec, Canada; [2]Deprtament of Microbiology and Plant Pathology, University of Pretoria, Pretoria, South Africa

E-mail: Franz.Lang@Umontreal.ca

Events in eukaryotic evolution such as the introduction of mitochondrial endosymbiosis, and the divergence of major lineages occurred a billion or more years ago. We perform phylogenomic analyses based on complete sets of mtDNA-encoded proteins, or on large collections of nucleus-encoded proteins. We are interested in the phylogenetic position of protists that are thought to branch near the animal-fungus divergence including *Monosiga*, *Capsaspora*, *Nuclearia*, *Amoebidium* and the Apusozoa.

AD9: Yan Zhou (UdeM)

Evaluation of the models handling heterotachy in phylogenetics inference

Recently KT has proposed a mixture branch length (MBL) model to handle heterotachy based on their simulation results that heterotachy has impaired the phylogenetic inference severely. However, the MBL model drastically increases the number of parameters due to the increased number of components in the mixture model. Until now, there are two heterotachous models available: the MBL model and the covarion model. Our comparisons using BIC and cross-validation show that the covarion model has a better fitness than the MBL model on all the real datasets so far we have analyzed. Our results further imply that the MBL model is expensive for the real heterotachous situation since the MBL assumes that heterotachous exists extensively along all the branches in the tree. Moreover, our results indicate that AIC over-estimate rich-parameter models and the cross-validation is rather a reliable method to evaluate models.

AD10 : Jean-Philippe Doyon (UdeM)

Inferring a duplication and speciation history from a gene tree

Auteurs: Cedric Chauve, Jean-Philippe Doyon, Nadia El-Mabrouk

We present a linear time and space algorithm for deciding if a gene tree can be obtained by a process involving only duplication and speciation events, and no gene loss. We show that in such a case, there is a single species tree that agrees with the gene tree. We show the results obtained on gene families of four yeast species.

AD11 : Yu Liu (UdeM)

A likelihood ratio test to identify fast-evolving sites in protein sequence, and its application on phylogenomic analysis

An artifact in phylogenetic analyses is Long Branch Attraction (LBA), which leads to the grouping of species with high evolutionary rates. In order to overcome LBA, one solution is to remove the potentially misleading data, most of which are characterized by rate heterogeneity. Here we present a Likelihood Ratio Test (LRT), which permits the progressive elimination of Highly Heterotachous sites that contribute to LBA. They are eliminated only in fast-evolving subgroups. Analysis of a published dataset shows that gradual removal of HH sites can efficiently decrease the effect of LBA.

AD12 : Jian Hong Chen (UdeM)

Sliding Between RNA Helical Elements Revealed by a Molecular Dynamics Simulation Study

The along-groove packing motif (AGPM) is found in sixteen places of ribosomal RNA. It consists of two double helices closely packed with each other via their minor grooves. At the center of the contact area, a GU base pair from one helix interacts with a Watson-Crick base pair of the other helix. An essential feature of AGPM is that it is able to bring together elements distant from each other in the secondary structure. This ability and the fact that AGPM has been found in many parts of the ribosome structure make this motif an essential element of the ribosome architecture. An important feature of the structure of AGPM is that it presumes the existence of particular dynamic characteristics. In particular, a shift of one helix along the other for one base pair does not affect the quality of packing. Therefore, two helices forming AGPM could slide one along the other without a necessity of overcoming a high energetic barrier. In this work, we used in silico approach for the analysis of the ability of two such helices to slide along each other. Molecular dynamics simulations were performed on multiple homogenous sequences packed in a pattern of along groove packing motif. They showed that some of sequence combinations can slide freely one helix upon another with the presence of minor constraints. This kind of conformational transition should be the intrinsic properties of helical RNA structures, and therefore might be of functional importance in folded RNA molecules such as in ribosomal structures.

AD13 : Tetsu Ishii

Selenocysteine tRNA (tRNAsec) delivers a rare amino acid selenocysteine in response to a UGA stop codon and a certain downstream stem loop signal in mRNA

It has an unusual secondary structure characterized by extended acceptor and D-stems. In E.coli, the acceptor stem is extended by one base-pair layer and the D-stem is extended by two base-pair layers. Structural modeling suggests that such an extension of these stems would drastically affect how a tRNA would interact with the ribosome. For instance, the extended acceptor stem alone would introduce a 2.8 angstrom length extension and a simultaneous rotation of 33 degrees about its axis. Therefore, if the acceptor stem is positioned in the standard peptidyl transferase center (PTC) the anticodon end of the tRNAsec would be misaligned with the mRNA codon. On the other hand, if the anticodon were to be positioned in the canonical way with the codon, the acceptor end would collide with the PTC. The D-stem extension can also affect the structure, as the canonical tertiary interactions 8-14-21 and 15-48 that provide for the rigidity in usual tRNAs are replaced by the two layer W-C extension of the D-stem. Such a loss would render a tRNA less rigid or more flexible. Based on the analysis of tRNAsec sequences obtained by in vivo selection and through computer modeling, we propose that the extension of the D-stem, which allows for a more flexible molecule, relieves the large affects caused by the long acceptor stem. In other words, we believe there is a structural compensation between the extended acceptor stem and the long D-stem in the tRNAsec.

AD14: François Boulay (UdeM)

Modeling tRNA conformity at the atomic level

Boulay F, Steinberg SV, Département de biochimie, Université de Montréal.

During translation, an aminoacylated tRNA (aa-tRNA) forms a complex with the elongation factor Tu (EF-Tu) for delivery to the ribosomal A site. The affinity of any aa-tRNA to Tu is composed of the affinities of the aminoacyl part and of the tRNA to Tu. For different amino acids and for different tRNA species, these latter affinities vary dramatically. Despite this, the affinities of cognate aa-tRNAs to Tu are relatively close to each other. The compensatory effect recently discovered by LaRiviere and colleagues (1) consists in the fact that in Nature, a stronger bound aminoacyl is usually combined with a weaker bound tRNA and vice versa. Such conformity constitutes a proofreading phenomenon, which allows non-cognate aa-tRNAs to be eliminated from translation. In this work, we used *in silico* approach to elucidate the nature of the aminoacyl-Tu interaction in the attempt to understand, which elements of the aminoacyl and in which way affect its affinity to Tu. By using energy minimization and molecular dynamics, we analyzed the roles played by different atoms and chemical groups of the aminoacyls in the interaction with EF-Tu. We found that the energy calculated for different complexes directly related the affinity of the complex measured experimentally (2). Using the minimal energy structures, we identified four residues of EF-Tu that are involved in the binding with the aminoacyl: H67, E226, T239 and N285. We present the effects of these four residues on different aminoacyls and their contribution to affinity.

1. LaRiviere FJ, Wolfson AD, Uhlenbeck OC, Science. 2001 Oct 5;294(5540):165-8.
2. Asahara H, Uhlenbeck OC, Biochemistry. 2005 Aug 23;44(33):11254-61.

APD1 : Romain Rivière (UdeM)

The building blocks of RNA molecules

We want to find a small set of motifs that cover RNA molecules. This problem is particulary interesting in the fields of motifs research and modelisation of large macromolecule, as it gives partial answers to the question of the existence of building blocks for RNA. We focused on the biggest crystallographic structure available today : the large subunit (LSU) of the ribosome of Haloarcula marismortui (pdb id 1JJ2).

We are going to construct a mapping between the nucleotides of a RNA molecule and the set of all possible motifs of given sizes of the given structure. We used the MC-annotate computer program to produce a graph of relations out of the LSU. We define a motif of this graph of relations as a subset of nucleotides that are connected by chemical interactions. We enumerated all possible motifs of the graph of relations with a tool developed in the lab and using state of the art subgraphs enumeration techniques. As those motifs are not easily usable in this form, we associated to each of the motifs a string that has the interesting property that two strings are equal if and only if the motifs are similar. This technique is known as canonical labeling. We used a self made modification of the world fastest canonical labeling program NAUTY to suit the particularity of RNA graph of relations. This allows us to construct the announced mapping.

With the given mapping, our problem is reduced to an algorithmically well known problem, named the hitting set problem. Unfortunately, this problem is known to be intractable. So, we used the best proved approximation algorithm that guaranty we would find a small set of motifs with regards to a minimum sized set. This idea behind this algorithms is quite simple. To construct a small set of motifs, we only need to always choose the one that cover the most number of remaining nucleotides.

The same ideas could be applied to find a small set of motifs with any given properties that cover not only the nucleotides but the most interactions of RNA molecules. We present results of building blocks for RNA in the form of small sets of graphs or cycles that cover most of the RNA macromolecules world. Some of them are well known RNA motifs.

APD2 : Ignacio Gonzalez Bravo

Providing expected values and confidence intervals for codon adaptation index: an application to oncogenic human papillomaviruses as a case study

Different organisms have differential preferences regarding codon usage. We present here an algorithm that calculates the expected Codon Adaptation Index (CAI) value for a set of sequences by generating random sequences with similar amino acid composition and similar total GC or GC3 than the input. It allows to identify deviations in CAI values as artefacts arising from compositional biases or as true outliers that might be biologically interpreted. The analysis of expected versus observed CAI values will be applied to oncogenic human papillomaviruses as example organisms.
