Colloque bio-informatique *Robert-Cedergren* Bioinformatics Colloquium

Université de Montréal
23-24 septembre 2004

Présentations orales et d'affiches
Poster and oral presentations

Programme

www.bioinfo.umontreal.ca/eve

Photo David To

BioneQ
Réseau québécois de bio-informatique

biT
Programme de bourses

Université de Montréal

Bienvenue au premier colloque bio-informatique Robert-Cedergren !

Ce colloque se veut le point de départ d'une tradition qui rassemblera, une fois l'an, la communauté universitaire oeuvrant en bio-informatique au Québec. L'objectif principal est de partager les derniers développements en ce domaine, de susciter une saine émulation par le biais d'un concours d'affiches et de présentations orales et enfin, de prendre en compte et de rendre compte de l'importance grandissante de la bio-informatique dans les sciences de la vie.

Les conférenciers invités seront Drs Christian Gautier, de l'Université Claude Bernard Lyon 1 et David Sankoff, de l'Université d'Ottawa.

Pour cette première édition, le programme de deux jours est divisé en deux sections :

- Étudiants du 2^e cycle : présentations orales – jour 1, affiches – jour 2
- Étudiants du 3^e cycle et stagiaires postdoctoraux : affiches – jour 1, présentations orales – jour 2

Cette année, dix-sept présentations orales et dix-neuf affiches seront en lice. 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
2^e cycle	2000 $	1000 $
3^e cycle	2000 $	1000 $
Postdoctorat	2000 $	1000 $

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

Gertraud Burger, Ph.D.

Welcome to the 1st annual Robert-Cedergren Bioinformatics Colloquium!

This first Colloquium is the starting point of an annual event gathering the academic community working in Bioinformatics in Quebec. The main purpose of this event is to share the latest Bioinformatics developments, to create a sound emulation by a friendly competition of posters and oral presentations to highlight the increasing role of Bioinformatics in life sciences.

Keynote speakers will be Dr. Christian Gautier, from Université Claude Bernard Lyon 1 and Dr. David Sankoff, from the University of Ottawa.

For this first edition, the two-day program is divided in two sections:

- Ms Students : oral presentations – Day 1, Posters – Day 2
- PhD Students & Postdoc : Posters – Day 1, Oral presentations – Day 2

This year, there will be 17 oral presentations and 19 posters. The university who has the best ranking will win the Robert-Cedergren Award.

This acrylic sculptur represents the principal tool of bioinformatician : the computer. The screen depicts 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	2000 $	1000 $
PhD	2000 $	1000 $
Postdocs	2000 $	1000 $

Enjoy this 1st annual Robert-Cedergren Bioinformatics Colloquium!

Gertraud Burger, Ph.D.

Comités/Committees

Présidente

Gertraud Burger, Université de Montréal

Comité d'organisation / Organizing Committee

Gertraud Burger, Université de Montréal
Marie Robichaud, programme de bourses biT
Sylvain Foisy, BioneQ

Arbitres / Referees

Franz Lang, Université de Montréal
Francois Major, Université de Montréal
Vladimir Makarenkov, Université du Québec à Montréal
Hervé Philippe, Université de Montréal
Pierre Rioux

Renseignements généraux

Accueil

L’accueil des participants se fera au Hall d’honneur du pavillon Roger-Gaudry, le jeudi 23 septembre dès 8 h 30. Les insignes d’identification vous seront remis au comptoir d’accueil.

Repas, pauses santé et cocktail

Les repas du midi, les pauses santé et le cocktail seront servis dans le Hall d’honneur.

General informations

Registration

The registration office is located in the Honor Hall in the Roger-Gaudry Building. Your identification badge will be available from 8:30 am, September 23.

Lunches, coffee breaks and cocktail

Lunches, coffee breaks and cocktail will be served in the Honor Hall.

Horaire des présentations orales (M-415)
Oral presentation schedule (M-415)

23 septembre 2004/September 23, 2004 (M.Sc.)

9h15 Mot de bienvenue/Opening remarks
Gertraud BURGER, Responsable, Programmes de bio-informatique
Alain CAILLÉ, Vice-Recteur à la recherche

9h30 Conférencier/Keynote Speaker
Christian GAUTIER, Université Claude Bernard Lyon 1

10h30 Présentation OM1/Presentation OM1
Arash SHABAN-NEJAD, Concordia University

11h00 Pause santé/Coffee break (Hall d'honneur, Honor hall)

11h30 Présentation OM2/Presentation OM2
Karine ST-ONGE, Université du Québec à Montréal

12h00 Présentation OM3/Presentation OM3
Jean-Eudes DUCHESNE, Université de Montréal

12h30 Dîner/Lunch (Hall d'honneur, Honor hall)

14h00 Présentation OM4/Presentation OM4
Geneviève BOUCHER, Université de Montréal

14h30 Présentation OM5/Presentation OM5
Marcos NAHMAD BENSUSAN, McGill University

15h00 Pause santé/Coffee break (Hall d'honneur, Honor hall)

15h30 Présentation OM6/Presentation OM6
Mathieu LAJOIE, Université de Montréal

16h00 Présentation OM7/Presentation OM7
Kossi LEPKOR, McGill University

16h30 Présentation OM8/Presentation OM8
Eric PAQUET, Université de Montréal

17h00 Présentation OD1/Presentation OD1
Paul DALLAIRE, Université de Montréal

17h30 Clôture/Closing

Horaire des présentations orales (M-415)
Oral presentation schedule (M-415)

24 septembre 2004/September 24, 2004 (Ph.D. & Postdoc.)

9h15	Mot de bienvenue/Opening Remarks *Michel BOUVIER, Directeur, Département de biochimie*
9h30	Présentation OD2/Presentation OD2 *Rachel BEVAN, McGill University*
10h00	Présentation OD3/Presentation OD3 *Nicolas RODRIGUE, Université de Montréal*
10h30	Présentation OD4/Presentation OD4 *Thomas LEPAGE, McGill University*
11h00	Pause santé/Coffee break (Hall d'honneur, Honor hall)
11h30	Présentation OD5/Presentation OD5 *Yaoqinq SHEN, Université de Montréal*
12h00	Présentation OD6/Presentation OD6 *Trevor BRUEN, McGill University*
12h30	Dîner/Lunch (Hall d'honneur, Honor hall)
14h00	Présentation OP1/Presentation OP1 *Annie CHÂTEAU, Université du Québec à Montréal*
14h30	Présentation OP2/Presentation OP2 *Amy HAUTH, Université de Montréal*
15h00	Pause santé/Coffee break (Hall d'honneur, Honor hall)
15h30	Présentation OP3/Presentation OP3 *Frédéric DELSUC, Université de Montréal*
16h00	Conférencier/Keynote speaker *David SANKOFF, Ottawa University*
17h00	Remise des prix/Awards (6)
17h15	Clôture/Closing
17h30	Cocktail (Hall d'honneur, Honor hall)

Horaire des affiches (Hall d’honneur)
Posters schedule (Honor hall)

23 septembre 2004/September 23, 2004 (Ph.D. & Postdoc.)

9h15 Mot de bienvenue/Opening remarks (M-415)
Gertraud BURGER, Responsable, Programmes de bio-informatique
Alain CAILLÉ, Vice-recteur à la recherche

9h30 Conférencier/Keynote speaker (M-415)
Christian GAUTIER, Université Claude Bernard Lyon 1

10h40 Affiche AD1/Poster AD1
Sivakumar KANNAN, Université de Montréal

11h00 Pause santé/Coffee break

11h30 Affiche AD2/Poster AD2
Yu LIU, Université de Montréal
Affiche AD3/Poster AD3
Charlotte HABEGGER-POLOMAT, Université Laval
Affiche AD4/Poster AD4
Naiara RODRIGUEZ-EZPELETA, Université de Montréal

12h30 Dîner/Lunch

14h00 Affiche AD5/Poster AD5
Béatrice ROURE, Université de Montréal
Affiche AD6/Poster AD6
Félix DOYON, Université de Montréal
Affiche AD7/Poster AD7
Brian CARRILLO, McGill University

15h00 Pause santé/Coffee break

15h30 Affiche AD8/Poster AD8
Tetsu ISHII, Université de Montréal
Affiche AD9/Poster AD9
Jianhong CHEN, Université de Montréal
Affiche AP1/Poster AP1
Amy HAUTH, Université de Montréal

17h00 Clôture/Closing

Horaire des affiches (Hall d'honneur)
Posters schedule (Honor hall)

24 septembre 2004/September 24, 2004

9h15	Mot de bienvenue/Opening Remarks (M-415) *Michel BOUVIER, Directeur, Département de biochimie*
9h30	Affiche AM1/Poster AM1 *Arash SHABAN-NEJAD, Concordia University* Affiche AM2/Poster AM2 *Valentin GUIGNON, Université de Montréal* Affiche AM3/Poster AM3 *Kush KAPILA, Concordia University* Affiche AM4/Poster AM4 *Geneviève BOUCHER, Université de Montréal* Affiche AM5/Poster AM5 *Philippe THIBAULT, Université de Montréal*
11h00	Pause santé/Coffee break
11h30	Affiche AM6/Poster AM6 *Claudia KLEINMAN, Université de Montréal* Affiche AM7/Poster AM7 *Mohamed TIKAH MARRAKCHI, Université de Montréal* Affiche AM8/Poster AM8 *Jean-François ST-PIERRE, Université de Montréal*
12h30	Dîner/lunch
14h00	Affiche AM9/Poster AM9 *Kossi LEPKOR, McGill University*
15h00	Pause santé/Coffee break
16h00	Conférencier/Keynote speaker (M-415) *David SANKOFF, Ottawa University*
17h00	Remise des prix/Awards (6) (M-415)
17h15	Clôture/Closing
17h30	Cocktail

PRÉSENTATIONS ORALES / ORAL PRESENTATIONS

OM1 : Arash SHABAN-NEJAD (Concordia)

FungalWeb Ontology: Designing a Formal Ontology of Fungal Genomics to analyse the large-scale enzyme-fungi interaction in OWL/DL Environment

Ontologies will play an important role in bioinformatics, as they do in other disciplines, where they will provide a shared source of precisely defined terms that can be communicated across people and applications. By using OWL (ontology web language)/DL(description logics) we try to highlight the features of the combination of a frame representation of OWL framework and expressive description logics. We show how to use Racer as DL reasoner to build and maintain sharable ontologies by revealing inconsistencies, hidden dependencies, redundancies, and misclassifications. Also, we suggest a way for gene regulation and protein secretion in our Fungal ontology.

OM2 : Karine ST-ONGE (UQAM)

Segment duplication and rearrangements are a major source of genomic diversity, both within a single organism, or when comparing different organisms. Formal models that define groups of duplicated and closely rearranged genes face a basic problem: in the most general cases, identifying them can require exponential computation time or space. Additional constraints must therefore be imposed in order to achieve efficiency, such as considering only one copy of each gene in each organism (Bergeron, Corteel, Raffinot, 2002), or restricting the comparison to only two organisms (He, Goldwasser, 2004).

We propose a pragmatic approach which reconciles biological relevance and computational efficiency. We also present some preliminary results on the comparison of bacterial genomes.

OM3 : Jean-Eudes DUCHESNE (UdeM)

L'étude de l'effet de différentes structures de données sur le temps d'exécution d'algorithmes de recherche de structures secondaires de l'ARN avec possibilité d'erreurs. Élaboration de règles générales pour l'extension de structures secondaires.

OM4 : Geneviève BOUCHER (UdeM)

Exploring the aggregation mechanisms of amyloid and amyloid-like oligomers.

Insoluble amyloid fibrils are found in several diseases such as Alzheimer's disease, type II diabetes and transmissible spongiform encephalopathies. Although the fibrillae observed in these diseases are structurally and histologically similar, the normally soluble proteins implied in their formation do not have, a priori, any common properties. Mounting evidence suggest that the toxicity observed in these diseases is not related to the fibrillae themselves, but to the soluble intermediate oligomers formed earlier in the process of fibrillogenesis. It is therefore of great interest to study and understand the mechanisms of oligomers formation. However, the experimental and detailed characterization of intermediate oligomers is complex: oligomers tend to be short-lived, they occur in very low concentration and are present under a high number of conformations and different degrees of aggregation. Considering these obstacles, in silico methods provide an interesting alternative approach that can complement efficiently experimental efforts. In particular, computer simulations should be able to provide reliable insights on the general properties associated with the aggregation mechanisms.

Even for small peptidic chains with an implicit solvent description, however, the aggregation process can take a time beyond the reach of standard simulation techniques such as molecular dynamics (MD) and Monte Carlo (MC), and alternative simulation techniques must be used. Using the activation-relaxation technique (ART nouveau) and an approximate free energy model with implicit solvent (OPEP), we study in details the mechanisms of aggregation some amyloid and amyloid-like peptides. Since ART events are defined directily in the energy landscape, the sampling efficiency is nearly insensible to the time scale, allowing the method to concentrate on the most relevant mechanisms. ART-OPEP has been tested extensively on a beta-hairpin [1] as well as a dimer and trimer of $A_{\beta16\text{-}22}$ [2], producing realistic folding or aggregation trajectories in agreement with experiments and standard simulations.

[1] G.H. Wei, Protein, 56(3):464-474(2004)
[2] S. Santini, Structure, 12(7):1245-1255(2004)

OM5 : Marcos Nahmad Bensusan (McGill)

In the Evolution of Cells towards Complexity and Why Life is as Diverse as it is: A Model of the Hox Gene Network Evolution

In our planet we are able to distinguish an astonishing diversity of life forms. A major problem in evolutionary developmental biology is to explain why life is as diverse as it is. If we want to accept the theory that all the existing creatures are derived from a common ancestor we must give an explanation of how different body plans arise and evolve. Phylogenetic analysis and fossil findings as well as high throughput gene expression data now available for many organisms provide clues to build phylogenetic trees and genetic networks of developmental processes.

The discovery of the homeotic genes, first in small organisms (Hashimoto (1941), Lewis (1978)) and later in vertebrates and mammals (Graham, et.al. (1989), Duboule and Dollé (1989)), has suggested that the body plans of several phyla is build on the same principle (Duboule, (1994)). Then, the evolution of development may occur by regulated mutations either upstream or downstream high conserved gene clusters (Wilkins (2002)).

However, a satisfactory explanation of why these precise developmental pathways have been drown by evolution still remains as an open paradigm. A major problem is how to relate the macromolecular level, where gene products interact and define the body plan of the organism, with the population level in which a population evolves towards a fitness increase due to natural selection and drift.

In this work, I will explore this problem under the light of a mathematical model, based in simple, but biologically reasonable hypothesis. There is theoretical and geographical evidence that speciation divergence has not occurred randomly. This event, referred as Cambrian explosion (Morris (1998)), marks the beginning of the Metazoan era, where most of the major phyla diverged from common ancestors. In particular, we observe that evolution gives origin to more complex organisms with improved abilities to regulate their physiological processes. Using Kauffman's model of genetic networks (Kauffman (1969), Kauffman (1993)), I employ the number of cell types to define the level of complexity in an organism (Solé, et. al. (2003)). This idea enables us to define an explicit algorithm of evolving networks that hopefully draws the pathways of cellular evolution and explain the direction in which evolution of organisms occurs.

Following a piecewise linear differential equations approach (Glass and Kauffman (1973)), for a specific gene network containing the homeotic (Hox) and segment polarity genes in the fruitfly Drosophila melanogaster, I will analyze how likely is the network to experiment a specific change on its topology, which modules or gene subsets tend to be conserved and which of these changes may be favored by evolution towards complexity.

OM6 : Mathieu LAJOIE (UdeM)

Un haplotype est une suite d'allèles associés à des sites polymorphes sur un segment chromosomique donné. L'étude des haplotypes est fondamentale en génétique médicale et en histoire des populations. Nous allons présenter un algorithme permettant de retracer l'histoire d'un haplotype à partir d'haplotypes ancestraux. L'idée est de trouver un ensemble minimal de recombinaisons et de conversions géniques permettant de créer l'haplotype en question à partir de ses haplotypes ancestraux.

OM7 : Kossi LEPKOR (McGill)

Conception d'un filtre variant dans le temps pour améliorer l'identification des protéines

En protéomique, l'identification des protéines à haut débit comprend l'acquisition des données par la spectrométrie de masse suivie de la recherche dans une base de données. Cependant, les spectromètres de masse sont corrompus par le bruit aléatoire et chimique qui induit des erreurs durant l'identification. Pour réduire l'effet de ces bruits, un filtre variant dans le temps a été développé. Cela implique la détermination de la composition fréquentielle des pics chromatographiques, la conception et l'implantation du filtre. L'application du filtre sur des données expérimentales a amélioré significativement l'identification des protéines de faible abondance.

OM8 : Eric PAQUET (UdeM)

Détermination de réseaux régulateurs par apprentissage machine

Le but de cette recherche est la détermination de la structure des réseaux régulateurs à partir d'analyses de résultats de biopuces (cf. microarrays). Nous développons actuellement une technique basée sur les réseaux Bayesiens qui permet d'utiliser conjointement des données séquentielles et non-séquentielles. La validité de notre approche est démontrée par l'application de notre nouvel algorithme sur des données simulées à partir de motifs de réseaux régulateurs réalistes.

****

OD1 : Philippe DALLAIRE (UdeM)

Hscan and hsxbl: computer tools for finding new microRNAs

Although only about 200 human microRNAs have been found, it is believed that many more exist. The goal of this project is to find novel human microRNAs using a bioinformatics approach: first by identifying the best candidate sequences in the human genome computationally, and second by validating them experimentally. First step: Pre-microRNAs form stem loop structures of roughly 70 to 100 nucleotides in length. In order to discover them from genomic sequence, we devised an approach that efficiently identifies thermodynamically sensible RNA hairpins whose sequence distribution share characteristics from the known human pre-miRNAs. We also devised a graph based approach to identify homologous hairpins from three different species. Three genomes (human, rat and mouse) were processed using these tools showing very acceptable running times and a human known pre-miRNAs enrichment as high as 51%. Second step: Low throughput biochemistry methods will be applied immediatley to test the most plaudible candidates, whereas a higher number of candidate sequences will be trialled using DNA chips in a further step.

OD2 : Rachel BEVAN (McGill)

A Fast and Accurate Method to Find Gene Rates

Overview: Knowledge of rate heterogeneity within combined data sets for phylogenetic analysis is important for accuracy in fitting models to the data. Although methods have been developed to allow for different model parameters based on different data partitions, there are currently no methods that provide the ability to compute the heterogeneity of the data a priori. A method to compute such relative rates of partitioned data (e.g. genes) will be presented. Simulation results reflecting the accuracy of the method will be given. Additionally the use of such rates in improving the maximum likelihood score, and bootstrap support of a tree, versus the concatenated data set will be demonstrated.

OD3 : Nicolas RODRIGUE (UdeM)

Reconstructing phylogenies from molecular data often rests on explicit mathematical models of DNA or protein evolution. Most models commonly applied today work under an important simplification: the assumption that evolutionary events at a particular site are independent from events at other sites. The main justification of this assumption is computational. This framework has allowed the development of efficient methods for determining the likelihood of a given phylogeny, whereby the likelihood is computed one site at a time, taking the product over all sites to obtain the overall likelihood (Felsenstein, 1981). It is in taking this product that one must assume that sites evolve independently, which is known to be incorrect in many cases. In the present work, a computational approach that allows one to consider the stochastic process underlying the evolution of a sequence as a whole is investigated. With this approach, functions that apply to the sequence as a whole can be considered directly in a phylogenetic inference. Here, a knowledge-based protein free energy function (Bastolla et al., 2001) was used with the idea that constraints related to the overall thermodynamic stability of a protein may be one of the major causes of interdependence. Using this approach, and a similar energy function, Robinson et al. (2003) have previously shown that biologically reasonable parameter estimates can be obtained when performing a statistical inference describing the relationship between two observed sequences. This work describes a generalization of the work of Robinson et al. from two taxa to n taxa, and introduces the approach in the context of phylogenetic inference.

References

Bastolla, U., Farwer, J., Knapp, E. W. & Vendruscolo, M. (2001). How to guarantee optimal stability for most representative structures in the Protein Data Bank. Proteins: Structure, Function, & Genetics, 44(2), 79-96.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Biology, 17:368-376.

Robinson, D. M., Jones, D. T., Kishino, H., Goldman, N., & Thorne, J. L., (2003). Protein Evolution with Dependence Among Codons Due to Tertiary Structure. Molecular Biology and Evolution, 20(10), 1692-1704.

OD4 : Thomas Lepage (McGill)

Modélisation du taux d'évolution en phylogénétique

L'horloge moléculaire fut longtemps une des hypothèses de base de l'analyse phylogénétique. Elle stipule que le taux d'évolution des séquences d'ADN est constant dans le temps. Toutefois, cette hypothèse est de plus en plus communément rejetée. Plusieurs équipes ont déjà essayé de déterminer une manière simple et efficace de décrire la variation du taux d'évolution dans un arbre phylogénétique (modèle covarion, modèle de Thorne et al.). Je présenterai ces travaux, avant d'exposer notre modèle, qui est à la fois plus réaliste biologiquement et moins approximatif, tout en ayant une vitesse de calcul comparable aux modèles antérieurs.

OD5 : Yaoqing Shen (UdeM)

Study of similar sequences from different organisms in Protist EST Program

Among the clusters in PEP database, there are a bunch of similar sequences from different organisms. Some of them haven't been annotated with any possible function. 42576 such clusters are picked out from 82936 sequences stored in PEP database. For these sequences from each organism, TBLASTX is done against sequences from all other organisms to find conserved sequences among different organisms. Simultaneously, BLASTN is done between the organisms to eliminate possible contaminations.

After removing possible contaminations, sequences with significant TBLASTX matches are classified according to the alignment length, identity between sequences and topology of the alignment. Twenty-nine pairs of sequences with an alignment length over 100 aa and identity over 80% are assembled by Phrap. Fifteen pairs of sequences are merged by Merger based on the topology of their alignments. For the elongated sequences, further investigation by BLASTX against nr database, domain search, and structure prediction based on sequence homology is done to find possible functional annotations**.**

OD6 : Trevor BRUEN (McGill)

Recent theoretical and experimental results concerning maximum parsimony

A general approach towards the maximum parsimony problem will be developed (based on the current notion of a compatibility graph), followed by an application of the general result towards a branch and bound algorithm. The results on simulated and actual data sets will also be presented.

****

OP1 : Annie CHATEAU (UQAM)

Les intervalles conservés sont un outil mathématique dédié à la bio-informatique. Récemment introduits, ils généralisent la notion d'adjacence conservée et permettent de comparer plusieurs génomes portant les mêmes gènes. Nous nous intéressons à l'application de la mesure de similarité offerte par les intervalles conservés dans le domaine de la reconstruction phylogénétique. Comment, par exemple, étiqueter les nœuds internes représentant les génomes ancestraux? Nous proposons dans cette présentation une méthode utilisant les propriétés formelles des ensembles d'intervalles conservés pour inférer des ensembles de candidats pour ces génomes ancestraux.

OP2 : Amy HAUTH (UdeM)

Sequence Self-Comparison: Using Tri-plots to Identify Direct and Inverted Repeats

Identification of sequence-similar regions are of interest to scientists as they can indicate recurring genomic elements (e.g. similar promoters), interspersed repeats (e.g. Alu sequences) and even gene duplications. Many bioinformatic tools exist to identify direct and inverted repeats both within a sequence and between sequences. Analysis of large sequences using these tools often identifies large quantities of repeats and the display and analysis of this data is often difficult.

I will describe visual analysis of dotplots for direct and inverted repeats within a sequence in the context of both contiguous and non-contiguous repetition. First, I describe analysis of a 5 kbp highly repetitive contiguous genomic region. Then, I detail analysis of non-contiguous repetition as it relates to comparison of mitochondrial genomes for two fast-evolving yeast variants. To facilitate visual analysis, I have created several expert-directed tools.

I conclude with a discussion of what needs to happen to automate analysis of tri-plots.

OP3 : Frederic DELSUC (UdeM)

Dating the evolutionary history of eukaryotes: does a relaxed clock reconcile molecules and fossils?

Molecular estimates of divergence times have often been controversial in generally appearing much more ancient than those suggested by the fossil record. The limited number of genes and species yet explored, and pervasive variations in evolutionary rates are the most likely sources of such discrepancies. Here we used concatenated amino acid sequences of 129 proteins for 36 eukaryotes allowing to determine the divergence times of several major clades, including animals, fungi, plants, and various protists. In order to handle significant variations of evolutionary rates and uncertainties of the fossil record, we used a Bayesian relaxed molecular clock simultaneously calibrated by 6 paleontological constraints. We estimated that, according to 95% credibility intervals, the eukaryotic kingdoms diversified 950-1,259 million years ago (Mya), animals diverged from choanoflagellates 761-957 Mya, and the debated age of the split between protostomes and deuterostomes occurred 642-761 Mya. These divergence times appeared to be robust with respect to prior assumptions of the method and paleontological calibrations. Interestingly, these relaxed clock time estimates are much more recent than those obtained under the assumption of a global molecular clock and thus appear closer to the ones suggested by fossils. However, the diversification of bilaterian animals still appear to be about 100 million years more ancient than the Cambrian boundary.

****

AFFICHES / POSTERS

AD1 : Sivakumar KANNAN (UdeM)

In typical genome projects, only ~50% of the protein-coding genes have been assigned to function, a fact that highlights the fundamental importance of gene identification methods in genomics research. It has become clear that an efficient and sensitive large-scale function annotation will require use of multiple features/attributes of the protein sequences, besides simple similarity. The main objective of this project is to develop a comprehensive analysis procedure to identify protein function using a machine learning method (predictive data mining). The goal is to detect hidden signatures and patterns in the integrated biological data, and to employ this new knowledge for deciphering genomic data at a large scale.

Predictive data mining is a search for patterns in an integrated data that can generalize and make rules, which when validated can be used for making accurate future decisions on new data. First, a set of known protein sequences is taken and described using various attributes that can be calculated directly from the sequence (e.g., physico-chemical properties). A data mining algorithm will then look for the patterns in the described data and learn rules from the observed patterns. Finally, the learnt rules will be tested and the best performing rules will be used for predicting the function of hypothetical proteins. An important objective of this project is to study the different ways of representing the sequences for efficient data mining. For proof of principle, an integrated set of well-curated mitochondrial data (GOBASE) is being mined exhaustively and an attempt to assign function to hypothetical proteins of a lobosean protist, Hartmanella vermiformis was made. The knowledge gained in this study will be applied to the massive data being generated in the Canadian Protist EST program (PEP).

AD2 : Yu LIU (UdeM)

Fungal Phylogeny Based on Ribosome Proteins

Fungi are a diverse group of organism, studied widely because of their commercial importance in biotechnology, agriculture and medicine, and because they provide simple model systems for illuminating the eukaryotic mode of life. Since the first fungal (also the first eukaryotes) Saccharomyces cerevisiae genome was completed in 1996, over the past few years the number of available, completely sequenced fungal genomes has increased from 1 to 10. Currently there are dozens of fungal genome and EST projects under way worldwide. This wealth of data has allowed a more complete understanding of fungal phylogeny. In this study, we use the genome and EST data from four principle division of fungi (Ascomycota, Basidiomycota, Zygomycota, Chytridiomycota) to address two unresolved issues in fungal phylogeny: 1) the exact placement of Archiascomycetes within Ascomycota. 2) Chytridiomycota is a monophylotic or paraphylotic group.

Ribosome genes are very conserved because of their significant role in the synthesis of protein. This makes them suitable to resolve deep phylogeny. In our preliminary analysis, 50 ribosome genes data were used to address fungal phylogeny problem. Bootstrap support for the internal branches among the fungi is high, and the four fungal divisions are clearly defined. *Schizosaccharomyces pombe*, the Archiascomycetes, is placed at the base of Ascomycota with very strong support. Chytridiomycota, which branch at the base of fungi, form a monophyly group with mediate support. This confirms with the analysis of the mitochondrial data.

AD3 : Charlotte HABEGGER-POLOMAT (Laval)

Modélisation de la glutamyl-ARNt synthétase d'*E. coli* par homologie

Notre projet de recherche allie l'enzymologie et la modélisation moléculaire dans l'étude de la relation entre structure et fonction lors de l'interaction entre une enzyme et un acide nucléique. De nombreuses questions concernant les interactions entre protéines et acides nucléiques restent à élucider. Pour y répondre, de nombreux chercheurs s'intéressent aux interactions entre les aminoacyl-ARNt synthétases (aaRS) et leurs ARN de transfert (ARNt) respectifs. Les interactions spécifiques entre aaRS et ARNt déterminent le code génétique et sont responsables de la fiabilité de la biosynthèse des protéines.

Au Laboratoire de Recherche sur la Biosynthèse des Protéines, nous étudions la glutamyl-ARNt synthétase (GluRS) d'*E. coli*, qui fixe le glutamate sur l'ARNtGlu lors de la biosynthèse des protéines. Une particularité du système GluRS-ARNtGlu est la nécessité de la présence de l'ARNt pour l'activation du glutamate, qui est la première étape de la réaction d'aminoacylation, ce qui fait de ce système d'étude un bon outil pour sonder les interactions entre protéines et ARN.

Un modèle des interactions entre la GluRS et l'ARNtGlu d'*E. coli* lors des premiers contacts a été proposé par Madore et al. (Biochemistry, 39(23):6791-8). Celui-ci suggère que le degré de flexibilité de l'extrémité 3' simple brin GCCA de l'ARNtGlu est un paramètre important dans la spécificité de l'interaction.

Dans ce modèle, un rôle important était suggéré pour la lysine 115, l'arginine 209 et l'arginine 48 de la GluRS. Nous avons donc exprimé et testé pour leur activité d'aminoacylation une première série de variants de la GluRS altérés au niveau de ces trois résidus. Il s'agit de K115A, K115W, R209A et R48A. Notre hypothèse de travail se voit renforcée par les résultats obtenus : le variant K115W montre une diminution d'activité de 60% par rapport à la GluRS sauvage, et les variants R209A et R48A perdent plus de 90% de leur activité d'aminoacylation. Ceci suggère que les trois résidus ciblés dans ces expériences ont un rôle déterminant dans la réaction d'aminoacylation. De plus, les études cinétiques montrent que c'est la constante catalytique kcat qui est affectée.

La structure tridimensionnelle de la GluRS d'*E. coli* n'a pas encore été résolue. Cependant, celle de son homogue chez Thermus thermophilus est connue et est présente dans la Protein Data Bank. Il est donc possible de modéliser par homologie la structure 3D de notre enzyme, puis de visualiser l'emplacement des résidus mutés dans sa séquence. Ceci nous permettra de mieux expliquer nos résultats et d'affiner notre stratégie de mutagenèse dirigée.

Ma présentation par affiche résumera les résultats obtenus en laboratoire lors de l'étude cinétique des variants de la GluRS d'*E. Coli* puis détaillera notre méthodologie et nos résultats préliminaires pour la modélisation, pour laquelle nous utilisons principalement Modeller.

AD4 : Naiara RODRIGUEZ-EZPELETA (UdeM)

Resolving the phylogenetic relationships between the three groups of primary photosynthetic organisms (red algae, green plants and glaucocystophytes) is of crucial importance to understand the origin of plastids. However, this issue has long been a subject of controversy, and no general agreement has been achieved so far. To clarify this question, we have sequenced 5,140 nuclear ESTs and the mitochondrial genome of Cyanophora paradoxa, a member of the glaucocystophytes, the less studied group of primary photosynthetic eukaryotes. Including this newly generated data, we have performed phylogenomic analyses based on nuclear, mitochondrial and plastid sequences and demonstrated that (i) the three groups of primary photosynthetic organisms form a robust monophyletic group with respect to other eukaryotic lineages and that (ii) all plastids form a monophyletic group with respect to cyanobacteria. Taken together, our results strongly support a single primary endosymbiosis at the origin of the three groups of primary photosynthetic eukaryotes.

AD5 : Béatrice ROURE (UdeM)

La majorité des modèles d'évolution supposent que tous les sites d'un alignement évoluent indépendamment selon le même mécanisme évolutif. Notre groupe a récemment développé un modèle bayésien qui prend en compte l'hétérogénéité des sites pour ce qui est des substitutions en acides aminés. Ce modèle, appelé CAT, est basé sur des profils de substitution définis par leur fréquence à l'équilibre pour les 20 acides aminés. Un profil particulier peut alors être affecté à chaque site. A partir d'un grand jeu de protéines bactériennes, la corrélation entre les profils générés par le modèle CAT et les données structurales est étudiée.

Par ailleurs, il a été montré que non seulement la vitesse d'évolution varie le long de la séquence, mais également au cours du temps pour une même position : cette variation temporelle est appelée hétérotachie. L'étude de l'affectation des profils dans différents groupes taxonomiques permet de mettre en évidence l'existence d'une hétérotachie qualitative correspondant à une variation en terme de profil de substitution pour un site donné.

Les problèmes de convergence des chaînes de Markov sur de très grands jeux de données seront également discutés.

AD6 : Félix DOYON (UdeM)

La structure canonique tridimensionnelle des ARNs de transfert (ARNt) a été élucidée voilà de ça déjà plusieurs années, mais l'importance relative des éléments la composant reste toujours un mystère. En effet, on ignore quels éléments peuvent être substitués sans que la structure et par conséquent la fonction de la molécule soient compromises. Nous avons comme objectif de caractériser les règles gouvernant la formation de structures tridimensionnelle d'ARN en utilisant comme modèle l'ARN de transfert pour sa relative simplicité. Une des régions les plus intéressantes chez les ARNt se situe au coin extérieur de la molécule, qui adopte une structure tridimensionnelle en forme de 'L'. À cet endroit (région DT), les boucles D et T forment de rigides interactions qui ont pour rôle de fixer correctement les deux domaines hélicoïdaux (D/Anticodon et T/Tige acceptrice) de la molécule l'un par rapport à l'autre, permettant ainsi à la molécule de jouer son rôle au niveau du ribosome et de la traduction. Afin d'étudier cette région-clé, nous avons construit une librairie combinatoire d'ARN fonctionnels *in vivo* où les boucles D et T ont été randomisées. Ceci nous a permis de séquencer un peu plus de 50 clones fonctionnels. L'analyse des séquences de ces 50 clones ainsi que des études *in silico* nous ont permis d'argumenter et de tirer des conclusions quant aux éléments cruciaux de la région DT. Par exemple, la paire de base reverse-Hoosgteen U54-A58, qui est présente dans la structure canonique, l'est aussi dans plus de la moitié des cas. Ceci n'est pas surprenant puisque celle-ci alloue le bon nombre de nucléotide dans l'appendice, ce qui assure une juxtaposition correcte des deux domaines hélicoïdaux. D'autre part, la région au-dessus de cette paire de base, qui au début ne semblait pas régie par des règles précises, a aussi démontré un caractère unique dans plus de la moitié des cas où cette paire de base était présente. Plus intéressant encore, lorsque la paire de base reverse-Hoogsteen n'est pas présente, la région DT adopte une autre conformation qui compense cette absence par un mécanisme que l'on pourrait bien retrouver chez les ARN de transfert mitochondriaux, qui sont reconnus pour leurs structures à tout le moins non-conventionnelles. Je présenterai donc trois modèles par lesquels la structure de la région DT peut être maintenue. Le premier se rapporte de façon assez étroite à la structure canonique des ARNts. Les deux autres constituent des structures totalement nouvelles qui n'ont jusqu'à maintenant jamais été observées chez les ARNts.

AD7 : Brian CARILLO (McGill)

The effect of the mass spectrometer's duty cycle was investigated by simulating the data directed acquisition behavior of the mass spectrometer for various duty cycles. The model was used to construct operating curves which dictate the trade-off between the quality of MS/MS spectra and the number of peptides fragmented. The optimal operating curve predicted significant differences in data quality between the various duty-cycles tested, indicating that significant improvements can be realized.

AD8 : Tetsu ISHII (UdeM)

The problem of how the nucleotide sequence of RNA molecules determines their tertiary structure and function remains essentially unsolved in spite of the considerable success achieved in the recent years in the X-ray crystallography and NMR-spectroscopy of different RNAs. Very little is known about the necessary and sufficient conditions for formation of the structure or for function of different RNA motifs and molecules. Even for such a relatively small and well-known molecule as transfer RNA (tRNA), these conditions are very far from being clearly elucidated. Thus, the existence of aberrant mitochondrial tRNAs, having secondary and tertiary structures very different from the standard and still able to deliver on their primary function, challenges all the rules and structural criteria for tRNA functionality established in the previous decades. Another unusual tRNA, selenocysteine-tRNA, has been found in all three major groups of organisms. This tRNA brings to the ribosome a rare amino acid selenocysteine in response to the UGA stop codon and a certain down-stream signal in mRNA. In all known cases, the structure of the selenocysteine tRNA is very different from the standard.

Moreover, it is different in eubacteria, archaebacteria and eukaryotes. The major subject of current project is the elucidation of the general constrains imposed on the nucleotide sequence and tertiary structure of this tRNA. For this, we will create combinatorial libraries of this tRNA in which certain positions in the nucleotide sequences would be randomized. These libraries will then be expressed in an E.coli strain in which the gene of the selenocysteine tRNA has been deleted. The clones capable of maintaining the normal incorporation of the selenocysteine into polypeptide chains will be selected and sequenced. The determined nucleotide sequences will be subjected to an intensive theoretical analysis, including molecular modeling, in order to elucidate common sequence and structural patterns. A comparative analysis of these patterns will shed new light on the mechanism of the selenocysteine incorporation and function of the selenocysteine tRNA.

AD9 : Jianhong CHEN (UdeM)

Ribosomes are ribonucleoprotein complexe that make proteins according to the genetic message contained in mRNA. The overall scheme of translation was determined about four decades ago, but a detailed mechanism of this process remained unclear mostly due to the absence of the structural data. The elucidation of the tertiary structure of the ribosome and its subunits has opened a new opportunities in understanding of the principles of the protein biosynthesis. It is known that during elongation cycle the ribosome undergoes specific conformational changes. To elucidate the mechanisms and functions of ribosome in this process, we decided to systematically compare the available crystal structures of subunits and to identify the essential different regions and common structural motifs. We made program suites for the comparison. And we found the most regions that were identified involves in one RNA helix gliding along another helix. These kinds of sliding don’t need extra energy and could go on smoothly. In order to get rid of the influences of irregular and inaccurate residues within RNA strands, we utilized standard form of RNAs to replace the real structure regions before the comparisons. Our final goal would be to draw a three-dimensional ribosomal picture indicating the flexible and solid regions and what to extent the movements are. By these, it would be possible for us to compare our analysis results with previous images of ribosome movement determined by cryoelectron microscopy (cryoEM). These findings allowed us to make some progress on the understanding of the mechanism of the ribosome dependent protein biosynthesis.

****

AP1 : Amy HAUTH (UdeM)

Analysis of structure and history of complex repetitive genome regions

Poster Authors : Amy M. Hauth, Gertraud Burger, B. Franz Lang

Some repetitive genome regions apparently result from amplification of a single ancestral sequence. For example, tandem repeats are regions which likely arose by a series of duplication events creating consecutive direct copies of the same sequence. Yet, genomic repeat elements occur in both direct and inverse orientation and regions undergo other changes such as mutation and recombination events. Our research studies these regions with a special interest on ones having complex combinations of segment similarity. We seek not only to characterize the current similarities and pattern structures within a region but also to determine the history of duplication, recombination and mutation events that could have led to its formation from a non-amplified ancestral sequence.

This poster presents our on-going research in this area. We show analysis of several complex repetitive regions in genomes: tandem repeats having complex pattern structures based solely on similarity in a direct orientation, contiguous regions containing numerous direct and inverse similarities and similarities between non-contiguous regions. In addition, we present the bioinformatic tools that enable these analyses (<A href=\"http://megasun.bch.umontreal.ca/People/ahauth/tools\">http://megasun.bch.umontreal.ca/People/ahauth/tools<A>).

****

AM1 : Arash Shaban-Nejad (Concordia)

FungalWeb Ontology: Design a Formal Ontology of Fungal Genomics to analysing the large-scale enzyme-fungi interaction in OWL/DL Environment

Ontologies will play an important role in the bioinformatics, as they do in other disciplines, where they will provide a shared source of precisely defined terms that can be communicated across people and applications. By using OWL (ontology web language)/DL (description logics) we try to highlight the features of the combination of a frame representation of OWL framework and expressive description logics. We show how to use Racer as DL reasoner to build and maintain sharable ontologies by revealing inconsistencies, hidden dependencies, redundancies, and misclassifications.

AM2 : Valentin Guignon (UdeM)

Les algorithmes exacts actuels de calcul de distance d'édition entre structures secondaires d'ARN ne sont pas assez rapides pour être employés sur de grosses bases de données pour des recherches fréquentes en temps raisonnable. Nous présenterons une revue des différents types d'éditions connues en tenant compte de leurs aspects biologiques. Ensuite, nous proposons d'élaborer un algorithme inspiré de ceux déjà existants mais offrant la possibilité d'être vectorisé pour améliorer la vitesse de calcul en tenant compte des types d'éditions présentés.

AM3 : Kush Kapila (Concordia)

PROTERAN: Protein Folding Trajectory Analysis Using Animated Terrain Evolution

In the last few decades there has been an explosion in the availability of biological data for the scientific community. The analysis of the data is crucial to understanding the underlying biology and answering vital questions. Various techniques such as data mining and clustering are being used on large data to extract useful information. Even sophisticated algorithms and analysis techniques will not be useful if the results are difficult to interpret or are not appropriately interpreted: thus visualization techniques play an important role in bridging this gap.

In my presentation, I will present PROTERAN, a novel 3D visualization technique which uses the metaphor of an evolving terrain to help identify the major states a protein folds into. I will demonstrate the program using clusters from beta-hairpin simulation data.

This work is in collaboration with Dr. Laxmi Parida and Dr. Ruhong Zhou, Computational Biology Center, IBM T.J. Watson Research Center. It forms part of my research for a Master's degree in the Computer Science and Software Engineering department of Concordia University being carried out under the supervision of Prof. Sudhir Mudur.

AM4 : Geneviève BOUCHER (UdeM)

Des agrégats de protéines normalement solubles sont retrouvés dans une vingtaine de maladies dont la maladie d’Alzheimer, la maladie de Parkinson, le diabète de type II et les encéphalopathies reliées aux prions. Malgré le fait que les fibrilles observées dans ces maladies partagent de grandes similarités structurales et histologiques, les protéines impliquées dans la formation de ces fibrilles n’ont, à priori, aucun lien entre elles, que ce soit au niveau de leur séquence ou de leur structure native. Un nombre croissant de résultats suggère que la toxicité observée dans les maladies reliées à la formation de fibrilles-amyloïdes ne serait pas liée aux fibrilles elles-mêmes, mais serait plutôt liée aux oligomères formés plus tôt dans le processus de fibrillogénèse. Cependant, la caractérisation expérimentale de ces intermédiaires est loin d’être simple puisque les oligomères ont une courte durée de vie et sont présents sous un grand nombre de conformations et de degrés d’agrégation différents, d’où l’intérêt d’étudier leurs mécanismes d’agrégation in silico.

Par contre, la formation des oligomères est un phénomène s’échelonnant sur une longue échelle de temps, ce qui limite le recours aux techniques de simulation numérique standard telles que la dynamique moléculaire et les simulations de Monte Carlo. En fait, peu de méthodes permettent d’étudier en profondeur tous les événements menant à l’agrégation des peptides et des méthodes alternatives doivent être utilisées.

Nous avons donc utilisé la méthode, récemment développée, d’activation-relaxation (ART-nouveau) combinée au potentiel d’énergie avec solvant implicite OPEP pour étudier les mécanismes d’agrégation d’un tétramère du peptide KFFE; le peptide de séquence KFFE étant le plus court peptide connu à former, *in vitro*, des fibrilles similaires à celles retrouvées dans les diverses pathologies. ART-OPEP a auparavant été utilisé efficacement pour replier une structure en épingle à cheveux bêta et pour étudier la formation de dimères et de trimères du peptide Abeta16-22.

Nous avons donc effectué près d'une vingtaine de simulations. Dans tous les cas, les peptides s’assemblent généralement en passant par des structures de basse énergie présentant les caractéristiques de feuillet bêta antiparallèles ou mixtes.

Nos résultats numériques présentent un processus d’assemblage caractérisé par l’ajout d’un peptide à la fois à une structure organisée déjà existante. Cependant, il est clair que le tétramère est trop petit pour produire une structure amyloïde stable et nos résultats ont montré que le tétrapeptide KFFE est trop court pour discriminer dynamiquement les différentes orientations à l'intérieur du feuillet bêta.

AM5 : Philippe THIBAULT (UdeM)

A constant time nucleic acid backbone construction algorithm was developed, and integrated to MC-Sym. RNA structures are built using two distinct processes: 1) the bases are positioned, and 2) the phosphodiester chain is constructed. This differs from the original MC-Sym, where the bases and the backbone were modeled together using a combinatorial discrete search. The original assumed no dependence between the relative orientation of two adjacent bases and their ribose conformation. This is clearly not the case, as using the original algorithm nearly 70% of the search space was incoherent. In the new, the ribose between two phosphate atoms is constructed using a numerical approximation method. X-ray crystallography and NMR ribose conformations were rebuilt with a precision of less than 1 Å of RMSD in constant time. Construction results compared to the original MC-Sym algorithm will be presented and discussed, including complete 3-D structures of the yeast tRNA-Phe.

AM6 : Claudia KLEINMAN (UdeM)

***In silico* approaches to RNA editing in plant mitochondria**

The term "RNA editing"'originally described a process of uridine insertion and deletion in mitochondrial transcripts of kinetoplastid protozoa. Over the years, the term has expanded to include several RNA re-tailoring processes that may occur post-transcriptionally or, in some cases, co-transcriptionally. Now, RNA editing is broadly defined as the modification of precursor RNAs to alter their sequences through insertion, deletion or specific substitution of nucleotides, so as to yield functional RNA species.

RNA editing is found in the nucleus, mitochondria and chloroplasts of a wide variety of organisms, for example viruses, protists, plants and animals, including humans. In plants, RNA editing has been identified as C to U as well as U to C conversions in both mitochondria and chloroplasts. In mitochondria, several hundred such changes are estimated to alter the coding text of the RNA population. Due to a lack of an in vitro editing system and to the difficulty to perform biochemical experiments, the key features of this process are still unknown.

Despite the biological importance of this phenomenon and the large body of biochemical data available, RNA editing has been only poorly investigated at a more formal level. We will use bioinformatics approaches to determine the signals conferring the specificity of this mechanism. We intend to organize existing data from diverse and dispersed sources such as sequence repositories, journal publications and web documentation. We will then use comparative genomics to find conserved elements in related species that could point to cis-elements of RNA editing. In addition, given the high frequency of editing in plant mitochondria, a large amount of sequence data of edited sites is available; intra-species comparisons will also be performed to use statistical methods to infer common elements.

AM7 : Mohamed TIKAH MARRAKCHI (UdeM)

Modélisation *in silico* de l'enzyme para-hydroxybenzoate hydroxylase (PHBH) et application au développement de catalyseurs artificiels

- Propriétés de l'enzyme para-hydroxybenzoate hydroxylase
  - Fonction et réaction enzymatique
  - Cinétique de la catalyse et stabilité (pH, Température)
  - Structure et ingénierie de l'enzyme PHBH
- Catalyseurs artificiels
  - Enzymes artificielles
  - Conception de novo de catalyseurs peptidiques
- Modélisation et étude *in silico* de l'enzyme PHBH
  - Études *in silico* déjà réalisées
  - Objectif de la recherche
  - Méthodologie

AM8 : Jean-François ST-PIERRE (UdeM)

1-Présentation d'une méthode d'étude de la dynamique des systèmes complexes sur des temps expérimentaux, la technique d'activation et de relaxation (TAR), ayant déjà servi à la simulation du repliement de courtes protéines.

2-Présentation d'une modification à la méthode, l'utilisation de coordonnées internes au lieu de coordonnées cartésiennes, qui permettrait une grande accélération à l'exécution de la méthode.

AM9 : Kossi LEPKOR (McGill)

Software-based filter design for de-noising mass spectra

In Proteomics, high throughput proteins identification involves mass spectrometry analysis and database search. The random and chemical noise that corrupts mass spectra leads to errors during peptide peak detection. To de-noise mass spectra, a time-varying filter was developed. This involved modeling spectral peak, determination of peak frequency content, design and implementation of the filter. The application of this filter to experimental spectra resulted in significant improvements in the signal-to-noise ratio.