COLLOQUE BIO-INFORMATIQUE
ROBERT CEDERGREN
BIOINFORMATICS COLLOQUIUM

Programme

Université de Montréal
8-9 novembre 2007

Présentations orales et affiches
Poster and oral presentations

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

Université de Montréal

Bienvenue au 4e 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 quatrième édition du colloque, les conférenciers invités sont :

Yves Van de Peer, Professeur en bio-informatique et génomique évolutionnaire, de l'Université de Gant en Belgique. Sa conférence a pour titre :

«The importance of gene and genome duplications for evolution and biological complexity: a case study on plants»

ainsi que

Sean R. Eddy, Chef du groupe Janelia Farm au Howard Hugues Medical Institute au Maryland, USA, dont la conférence est intitulée :

« The modern RNA world: relics, regulators, and rogues»

Au total, dix-sept présentations orales et quinze affiches seront en lice dans quatre catégories.

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 4th annual Robert Cedergren Bioinformatics Colloquium!

This fourth 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 Yves Van de Peer, Professor, Bioinformatics and Evolutionary Genomics, Gent University, Belgium and Sean R. Eddy, Group Leader, Janelia Farm, Howard Hugues Medical Institute, Maryland, USA.

This year, 17 oral presentations and 15 posters will compete in 4 categories.

Individual awards will be given in the following categories:

	Best oral presentations	Best posters
MSc	1000 $	500 $
Ph.D.	1000 $	500 $

Enjoy this 4th annual Robert-Cedergren Bioinformatics Colloquium!

Gertraud Burger, Ph.D.
Leader
Bioinformatics graduate programs

Comités / Committees

Comité d'organisation / Organizing Committee

Gertraud Burger
Marie Robichaud
Elaine Meunier

Arbitres / Referees

Philip Awadalla (CHU Ste-Justine)
Nadia El-Mabrouk (UdeM)
Sylvie Hamel (UdeM)
Damian Labuda (CHU Ste-Justine)
Franz Lang (UdeM)
Sébastien Lemieux (UdeM)
Sabin Lessard (UdeM)
Vladimir Makarenkov (UQAM)
Alejandro Murua (UdeM)
Hervé Philippe (UdeM)
Reza Salavati (McGill)
Elisabeth Tillier (U. Toronto)
Marcel Turcotte (U. Ottawa)

Responsables de séance / Session Chairs

Nadia El-Mabrouk
Franz Lang
Sébastien Lemieux
Marcel Turcotte

Renseignements généraux / General information

Accueil / Registration

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

Prenez note que le Colloque aura lieu:

le 8 novembre dans le M-415 et le Hall d'Honneur, pav. Roger-Gaudry
le 9 novembre dans le Z-110 et le corridor adjacent (Z-100), pav. Claire-McNicoll

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 8.

Please remind that the Colloquium willl take place:

November 8, in the Room M-415, and the Honor Hall, Roger-Gaudry Bldg
November 9, in the Room Z-110, and the near area (Z-100), Claire McNicoll Bldg

Pauses santé et cocktail / Coffee breaks and cocktail

Les pauses santé et le lunch du 8 novembre seront servis dans le Hall d'honneur.
Les pauses santé, le lunch et le cocktail du 9 novembre seront servis dans le corridor Z-100.

Coffee breaks and the lunch will be served in the Honor Hall, for November 8 only.
Coffee breaks, lunch and cocktail will be served in the Z-100 area for November 9.

HORAIRE
Jeudi, 8 novembre 2007

9 h 15 **Ouverture du colloque : Pierre Boyle, vice-doyen à la recherche, Faculté de Médecine, Université de Montréal**

9 h 30 ***Conférence 1 : Yves Van de Peer, Gant University, Belgium***
«The importance of gene and genome duplications for evolution and biological complexity: a case study on plants»

10 h 30 Pause santé

11 h 00 **OD1 : Béatrice Roure (Université de Montréal)**
The probability of identical substitution profiles as a criterion to detect positions involved in a functional shift

AM1 : Philippe Nadeau (Université de Montréal)
Étude par simulation des signatures génomiques de la sélection naturelle

AM2 : Martin Smith (Université de Montréal/Université Laval)
Evolution, Characterization and Genomic Distribution of two Groups of SIDER in Leishmania Protists

AM3 : Julie Hussin (Université de Montréal)
Recombinaison méiotique et structure du génome humain

11 h 30 **OD2 : Mathieu Lajoie (Université de Montréal)**
Evolution of tandemly arrayed genes in multiple species

AM4: Xiaoquan Yao (University of Ottawa)
Functional divergence in methionine aminopeptidase (MAP) among human, yeast and E. coli: a bioinformatic approach

AM5: Diala Abd Rabbo (Université de Montréal)
Identification des profils d'expression associés à une mutation de l'un des gènes BRCA1 et BRCA2 dans les cellules non-tumorales de l'épithélium de surface de l'ovaire

AM6 : Sam Khalouei (University of Toronto)
Translation initiation in human immunodeficiency virus type 1:
analysis of HIV-1 5'-untranslated region

OM = Oral M.Sc. OD = Oral Ph.D. OPD = Oral Postdoc	AM = Affiche M.Sc. Poster AD = Affiche Ph.D. Poster AHC = Affiche Hors concours
Oral presentations : M-415	**Posters : Hall d'honneur**

12 h 00 **OD3 : Huiling Xiong (University of Ottawa)**
Using Generalized Procrustes Analysis (GPA) for normalization of cDNA microarray data

AM7 : Marie Pier Scott-Boyer (Université de Montréal)
Computational Annotation of Non-Coding RNAs in Candida albicans

12 h 30 Lunch + Posters session M.Sc.

13 h 30 **OD4: Yaoqing Shen (Université de Montréal)**
"Unite and conquer": enhanced prediction of protein subcellular localization by integrating multiple specialized tools

AHC1 : Elizabeth Tillier (University of Toronto)
A Fast and Flexible Approach to Oligonucleotide Probe Design for Genomes and Gene Families

14 h 00 **OD5 : Dapeng Zhang (University of Ottawa)**
Functional insight into maelstrom proteins in the germline small RNA pathway: a novel domain with a derived DnaQ 3'-5' exonuclease fold and its lineage-specific evolutionary expansion/loss

14 h 30 **OD6: Sivakumar Kannan (Université de Montréal)**
Unassigned murf1 of kinetoplastids codes for NAD2

15 h 00 Pause santé

15 h 30 **OD7 : Alireza Shaneh (McGill University)**
On Acid-Base Complementarity in Protein Interaction Networks

16 h 00 **OD8 : Karine Cyrenne (Université de Montréal)**
Modelisation of RNAs tridimensional structures

16 h 30 **OD9 : Mathieu Coinçon (Université de Montréal)**
Au delà de la modélisation: Étude dynamique de la catalyse chez les fructose-1,6-bisphosphate aldolases

17 h 00 **OD10 : Lilianne Dupuis (Université de Montréa)**
Multiscale simulations of protein flexibility

OM = Oral M.Sc. OD = Oral Ph.D. OPD = Oral Postdoc	AM = Affiche M.Sc. Poster AD = Affiche Ph.D. Poster AHC = Affiche Hors concours
Oral presentations : M-415	**Posters : Hall d'honneur**

HORAIRE
Vendredi, 9 novembre 2007

9 h 00 ***Conférence 2 : Sean R. Eddy, Howard Hugues Medical Institute, MD, USA***
« The modern RNA world: relics, regulators, and rogues»

10 h 00 **OM1: Marie Pier Scott-Boyer (Université de Montréal)**
Computational Annotation of Non-Coding RNAs in Candida albicans

AD1 : Nicolas Rodrigue (Université de Montréal)
A mechanistic account of amino acid of codon preferences in models of protein-coding nucleotide sequence evolution

AD2 : Yan Zhou (Université de Montréal)
Handling heterotachy with MBL model and Covarion model

AD3: Sivakumar Kannan (Université de Montréal)
Evaluating ORF Function Predictions using Domain-specific Knowledge

10 h 30 **OM2 : Sam Khalouei (University of Toronto)**
Translation initiation in human immunodeficiency virus type 1: analysis of HIV-1 5'-untranslated region

AD4: Hamed Shateri Najafabadi (McGill)
Prediction of protein interaction maps for Trypanosoma brucei and Trypanosoma cruzi: all for one, one for all

AD5 : Yaoqing Shen (Université de Montréal)
"Unite and conquer": enhanced prediction of protein subcellular localization by integrating multiple specialized tools

AD6 : Claude Bhérer (Université de Montréal)
Contribution génétique différentielle des fondateurs aux pools géniques régionaux du Québec

11 h 00 **Pause-santé**

11 h30 **OM3: Mathieu Courcelles (Université de Montréal)**
MSdatabase: a novel proteomics and phosphoproteomics analysis platform

AD7: Lilianne Dupuis (Université de Montréal)
Multiscale simulations of protein flexibility

AD8: André Levasseur (Université de Montréal)
Inférence d'orthologie par contexte génomique

OM = Oral M.Sc. OD = Oral Ph.D. OPD = Oral Postdoc	AM = Affiche M.Sc. Poster AD = Affiche Ph.D. Poster AHC = Affiche Hors concours
Oral presentations : Z-110	**Posters : Corridor Z-100**

12 h 00 **OM4 : Ziyu Song (University of Ottawa)**
The mechanism and usage of the "SD-independent" translation initiation pathway in the prokaryotes

AHC2 : Alejandro Murua (Université de Montréal)
Modeling replicated time-course evolution of gene-expression profiles

12 h 30 Dîner + Posters session

13 h 30 **OM5 : Julie Hussin (Université de Montréal)**
Signatures génomiques de sélection et structure des populations humaines

AHC3: TBestDB: a taxonomically broad EST database

14 h 00 **OM6 : Martin Smith (Université de Montréal / Université Laval)**
Evolution, Characterization and Genomic Distribution of two Groups of SIDER in Leishmania Protists

14 h 30 **OM7 : Wafae El Alaoui**
Datation moléculaire et artefact dû à la saturation mutationnelle

15 h 00 Pause

15 h 30 **OPD1 : Christian Landry (Université de Montréal)**
An in vivo map of the yeast interactome

16 h 00 **OPD2 : Véronique Ladret (Université de Montréal)**
Statistical properties of allelic haplotype classes and statistical tests of natural selection

16 h 30 Remise des prix : Gertraud Burger

16 h 45 Mot de clôture : Muriel Aubry

17 h 00 Cocktail et bouchées

OM = Oral M.Sc. OD = Oral Ph.D. OPD = Oral Postdoc	AM = Affiche M.Sc. Poster AD = Affiche Ph.D. Poster AHC = Affiche Hors concours
Oral presentations : Z-110	**Posters : Corridor Z-100**

CONFÉRENCES

Yves Van de Peer

Bioinformatics and Evolutionary Genomics, Gent University, Belgium

The importance of gene and genome duplications for evolution and biological complexity: a case study on plants

Recent analyses of eukaryotic genome sequences have revealed that gene duplication, by which identical copies of genes are created within a single genome by unequal crossing over, reverse transcription, or the duplication of entire genomes, has been rampant. The creation of extra genes by such duplication events has now been generally accepted as crucial for evolution and of major importance for adaptive radiations of species and the general increase of genetic and biological complexity. We have developed software to identify remnants of large-scale gene duplication events and more recently, we have also developed mathematical models that simulate the birth and death of genes based on observed age distributions of duplicated genes, considering both small and large scale duplication events. Applying our model to the model plant Arabidopsis shows that much of the genetic material in extant plants, i.e., about 60% has been created by several ancient genome duplication events. More importantly, it seems that a major fraction of that material could have been retained only because it was created through large-scale gene duplication events. In particular transcription factors, signal transducers, and regulatory genes in general seem to have been retained subsequent to large-scale gene duplication events. Since the divergence of (duplicated) regulatory genes is being considered necessary to bring about phenotypic variation and increase in biological complexity, it is indeed tempting to conclude that such large scale gene duplication events have indeed been of major importance for evolution. By using microarray expression data, we also show that the mode of duplication, the function of the genes involved, and the time since duplication play important roles in the divergence of gene expression and therefore in the functional divergence of genes after duplication. We have studied the expression divergence of genes that were created during large and small-scale gene duplication events by means of microarray data and have investigated both the influence of the origin and the function of the duplicated genes on expression divergence. We found that duplicates that have been created by large-scale duplication events and that can still be found in duplicated segments have expression patterns that are much more correlated than those that were created by small-scale duplications or those that no longer lie in duplicated segments. Moreover, the former tend to have highly redundant or overlapping expression patterns and are mostly expressed in the same tissues, while the latter show asymmetric divergence. In addition, a strong bias in divergence of gene expression was observed towards gene function and the biological process genes are involved in. For instance, genes involved in signal transduction and response to stress and external stimulus have expression patterns that diverged quickly after duplication. In contrast, genes involved in conserved processes such as cell organization diverge more slowly after duplication.

Sean R. Eddy

Group Leader, Janelia Farm, Howard Hugues Medical Institute, Maryland, USA

The modern RNA world: relics, regulators, and rogues

The discoveries of microRNAs, riboswitches, and numerous bacterial regulatory RNAs have created a new wave of interest in the diversity of noncoding RNA structure and function. A number of genome-wide screens for novel functional noncoding RNAs have been carried out by my lab and by others, using both computational and experimental methods. Two key questions in this field remain largely unanswered. First, in a large-scale screen, how can we distinguish a functional noncoding RNA from a messenger RNA, when coding regions as small as six amino acids are known? Second, how can we distinguish functional from nonfunctional noncoding RNA transcripts - how much transcriptional noise do organisms tolerate? These questions are of central importance, because transcriptomics data is interpreted by some as evidence for a large amount of functional noncoding RNA transcription, but this interpretation remains poorly supported.

PRÉSENTATIONS ORALES / ORAL PRESENTATIONS

Oral M.Sc.

OM1 : Marie Pier Scott-boyer (UdeM)

COMPUTATIONAL ANNOTATION OF NON-CODING RNAS IN CANDIDA ALBICANS

Candida albicans is a fungal pathogen causing systemic candidiasis, an important cause of mortality in immunocompromised patients. The annotation of the genome was performed with the exception of non-coding RNAs (ncRNAs). ncRNAs are involved in several essential processes in eukaryote cells for example: tRNA, snoRNA and rRNA. Antifungal drug discovery is limited by the availability of suitable drug targets and identifying essential ncRNAs would increase the pool of possible targets. ncRNAs are difficult to identify from genome sequence alone because secondary structure rather than sequence is responsible for their function. We scanned C. albicans haploid genome with the tools RFAM/Infernal [1], QRNA [2], RNAz [3] and Dynalign [4]. We will present results concerning the families identified by the different methods and the threshold scores used to determined significant hits. We will finally discuss the multiple strategies to prioritize predictions for validation. Known ncRNAs have mostly been identified in higher eukaryotes. Thus, to have optimal drug targets against fungi, we need a method to predict ncRNAs without using homology to a related organism or searching for a defined structure. To solve this problem, we propose an approach that enumerates (from genomic sequence) all possible compact motifs involving 1 or 2 base pairs that could form within all possible secondary structures. We then evaluate the ability to discriminate between structured and non-structured RNA sequences by comparing the distributions of enumerated motifs. After analyzing the results, we observed that most of the discriminating signal seems to come from a difference in dinucleotide composition in ncRNAs. We are currently developing an approach to use this information combined with a machine learning system (HMM) to recognize patterns present in the upstream region of genes to improve the prediction of ncRNAs. Preliminary results will be presented.

[1] Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy S R and Bateman A. Nucleic Acids Res., 33:D121-D124 (2005)
[2] Rivas E and Eddy S R. BMC Bioinformatics, 2:8 (2001)
[3] Washietl S, Hofacker I L and Stadler P F. Proc. Natl. Acad. Sci. U.S.A., 102:2454-2459 (2005)
[4] Mathews D H and Turner D H. J. Mol. Biol., 317:191 (2002)

OM2 : Sam Khalouei (UToronto)

TRANSLATION INITIATION IN HUMAN IMMUNODEFICIENCY VIRUS TYPE 1: ANALYSIS OF HIV-1 5'-UNTRANSLATED REGION

The human immunodeficiency virus type 1 (HIV-1) is dependent on the host transcription and translation machinery to complete its life cycle. Different hypotheses have been postulated regarding the mechanism of translation initiation in HIV-1. It has been shown that HIV-1 mRNAs undergo cap-dependent ribosomal scanning and viral gene expression is modulated through varying signal strengths of the Kozak consensus sequences. Recent reports have shown evidence of cap-independent translation initiation mechanisms in HIV-1, such as the direct binding of ribosome at internal ribosome entry sites (IRES), distant from the 5' cap. The cap-dependent ribosomal scanning hypothesis predicts that there should be a selective pressure against ATG usage in optimal context in the HIV-1 5'UTR to avoid their erroneous selection by the scanning ribosome which hampers the detection of the true downstream translation initiation codon. The IRES-dependent translation initiation hypothesis, on the other hand, predicts that there is no such selective pressure since the ribosome directly binds at an internal site without having to scan the sequence. We evaluated these two hypotheses based on their prediction regarding selective pressure against ATG usage in optimal context in the HIV-1 5'UTR. Our results show that there is indeed a selective pressure against such ATG's, which supports the cap-dependent translation initiation and contradicts the IRES-dependent translation initiation hypothesis. In the second part of our study we used a bioinformatics approach, including position weight matrix and perceptron, to analyze a region consisting of the 50 nucleotides upstream of the translation initiation codon in 331 unique HIV-1 and 24850 unique human

sequences in an attempt to find conserved sites specific to HIV-1. We found that sites -17 and -25 (relative to the translation initiation codon) are highly conserved. Furthermore, using perceptron, we were able to separate HIV-1 and human sequences based on these upstream 5'UTR 50-mers. The strong site conservation specific to HIV-1 5'UTR 50-mers points to the possibility of as yet unknown cap-independent translation initiation mechanisms in HIV-1. The clear separation of HIV-1 and human sequences based on these 50-mers also raises the promising possibility of designing novel antiviral drugs that specifically suppress translation initiation in HIV-1 with little effect on human translation.

OM3 : Mathieu Courcelles (UdeM)

MSDATABASE: A NOVEL PROTEOMICS AND PHOSPHOPROTEOMICS ANALYSIS PLATFORM

Over the past five years, remarkable advances in high resolution mass spectrometry and affinity media have facilitated large-scale phosphoproteome analyses in support to molecular and cellular biology projects. The availability of DNA sequences information of complete genomes combined with the development of informatics tools to automate the interpretation of MS/MS spectra and to locate phosphorylation sites now enable high-throughput phosphoproteomics analyses. These developments had significant impacts on the study of global cell signaling events in response to chemical stimulation and on the design of specific kinase inhibitors for cancer drug therapies. In view of the wealth of information generated by these large scale phosphoproteomics experiments, novel and improved bioinformatics tools are now require to profile changes in phosphorylation and associate interacting partners involved in specific signaling cascade events. To this end, we developed a bioinformatics platform tailored to address the pressing needs of phosphoproteomics analyses. The platform was conceived to store into a relational database all phosphorylation site identifications to collect all evidences from LC-MS/MS experiments and to evaluate false positive identifications. Known and characterized phosphorylation sites (in vivo function, kinase) from Phospho.ELM, the most comprehensive phosphorylation database, and Swiss-Prot are incorporated to conveniently track previous records. Many existing tools have been interfaced with the platform for specific phosphoproteome analysis. PPSP, a phosphorylation sites predictor, is used to identify potential kinases for identified sites. Overrepresented phosphorylation motifs in the dataset are searched with Motif-X. Motifs from uncharacterized kinases can be discovered from this approach. Finally, to gain further knowledge and have a global overview of phosphoproteome, the dataset of phosphoproteins are mapped on protein-protein interaction network from cPath database using Cytoscape. Similarly for molecular interaction and reaction networks, KEGG pathways are mapped with GenMapp. The application of these tools will be demonstrated for the identification of new MAP kinase substrates in stimulated IEC6 rat cells and in the study of specific signaling pathways activated during the stimulation of J774 mouse macrophage cells with interferon-γ.

OM4 : Ziyu Song (UOttawa)

THE MECHANISM AND USAGE OF THE "SD-INDEPENDENT" TRANSLATION INITIATION PATHWAY IN THE PROKARYOTES

In the prokaryotes, it is widely accepted that the translation initiation involves a direct recognition of the initiation region on the mRNA by the 30srRNA mediated by the base paring of Shine-Dalgarno (SD) motif in the 5'-untranslated region (5'UTR) of mRNA and anti-SD motif in 3'-end of 16srRNA. However, a significant percentage of genes, which lacked of SD motifs in their 5'UTR or even lacked of the entire or most 5'UTR were found in the genome of some prokaryotes, especially in some crenarchaeota archaea. These genes without SD motifs must adopt an alternative "SD-independent" translation initiation pathway. A prevalent hypothesis for the "SD-independent" pathway is that 30srRNA, first forming a 30S-fMet-tRNAi complex with initiation tRNA, can recognize the initiation site of mRNA and thus activate the translation initiation. This is similar to the "scanning" mechanism of translation initiation in the eukaryotes. Recently, a"70srRNA hypothesis", which is a novel pathway proposed for translation initiation of genes without SD, argues that 70srRNA can bind to the leaderless genes and start the translation without the helps of any initiation factors and initiation tRNAs. The former researches on these two hypothesis were all conducted in the lab using the "in vitro" system. However, no clear evidence was found to support them in the "in vivo" system. One task of my research is to test these two hypotheses by analyzing the translation initiation sites with bioinformatics methods. The results of my research strongly supported the "70srRNA" novel pathway. In addition, a hypothesis that explains the different usage of "SD-dependent" pathway and "SD-independent" pathway states that the "SD-dependent" pathway is mostly used

in the internal genes or last genes in polycistronic transcripts while the "SD-independent" pathway is mostly used in the single genes in monocistronic transcripts or first genes in polycistronic transcripts. However, the methods of the researches supporting this hypothesis are problematic and limited only in the crenarchaeota archaea. Based on this, the other task of my research is to use a more convincing way to examine the different usages of the "SD-dependent" pathway and the "SD-independent" pathway and to verify if this is universal in all the prokaryotes. The current results of my research are supportive.

OM5 : Julie Hussin (UdeM)

SIGNATURES GÉNOMIQUES DE SÉLECTION ET STRUCTURE DES POPULATIONS HUMAINES

La variabilité génétique chez l'humain est causée par les processus de mutation, qui crée des sites polymorphes (SNPs), et de recombinaison, qui crée de nouveaux haplotypes, i.e. de nouvelles combinaisons d'allèles pour les SNPs adjacents. Cette diversité génétique est influencée par l'histoire démographique des populations (croissance, migration), par la dérive génétique ainsi que par la sélection naturelle. Depuis son apparition en Afrique, l'homme moderne s'est adapté aux nombreux changements, qui ont donné lieu à des pressions sélectives façonnant le génome humain pour, par exemple, favoriser les phénotypes bénéfiques. Comprendre les éléments génétiques correspondants, cibles de la sélection, nous informe sur les évènements qui ont profilé l'humain. L'identification de loci soumis à la sélection naturelle est un moyen intéressant de trouver des variants génétiques fonctionnels importants. Dans le cas de la sélection positive adaptative, la prévalence des allèles sélectionnés augmente dans la population, ce qui marque les séquences d'ADN de "signatures" de sélection. Ces empreintes peuvent être détectées par comparaison avec la variation génétique neutre des séquences. Cependant, déterminer si une signature identifiée est due à la sélection ou aux effets de l'histoire démographique est un grand défi. Pour le relever, depuis plus de 20 ans un grand nombre de tests statistiques ont été développés. Ils utilisent les données de polymorphisme et se basent sur les modèles de mutations existants (Infinite site model, Infinite allele model). Une nouvelle statistique, Sv2, a été développée par notre groupe, et découle des classes alléliques d'haplotypes, qui combine l'information issues des deux modèles. Chaque classe regroupe les haplotypes partageant le même nombre m de nouvelles mutations. Ainsi, la classe allélique des haplotypes Am=2n regroupe les haplotypes avec exactement 2 allèles dérivés. Dans le cadre d'un projet commun afin de valider Sv2 comme indice de sélection, mon travail consiste à caractériser Sv2 sur les données empiriques dans différentes populations humaines. Grâce à l'abondance de données sur la variation haplotypique des populations humaines, la détection de régions sous sélection se fie grandement sur l'utilisation de données empiriques. Je présenterai les données disponibles et les biais qui peuvent être identifiés, corrigés ou évités. Par une approche par sliding window, j'ai développé un outil nous permettant de calculer les distributions empiriques de différentes statistiques utilisées fréquemment : test D (1), test H (2), indice moyen de diversité génétique FST, test LRH (3), test iHS (4). Mon objectif est de comparer ces distributions à celle de Sv2 et de comprendre comment ces différentes informations nous aiderons à identifier des régions précises sous sélection naturelle et leur mode de sélection (positive, balancée, etc.). Ces résultats seront comparés à la distribution de Sv2 calculée sous différentes conditions, pré-établies par simulations (travail par Philippe Nadeau). Je présenterai des résultats préliminaires sur le comportement de Sv2, obtenus par l'analyse des données ENCODE sur trois grandes populations humaines (africaine, asiatique et européenne). Finalement, pour mieux comprendre la structure du génome, mon travail futur consistera à m'intéresser aux liens qui existent entre les patrons de sélection et de recombinaison.

[1] Tajima F., 1989
[2] Fay J. and Wu C., 2000
[3] Sabeti P. C. et al., 2002
[4] Voight et al. 2006

OM6 : Martin Smith (UdeM/ULaval)

EVOLUTION, CHARACTERIZATION AND GENOMIC DISTRIBUTION OF TWO GROUPS OF SIDER IN LEISHMANIA PROTISTS

A large family of extinct retroposons known as SIDERs (Short Interspersed Degenerated Retroposons) has recently been identified in the genome of the parasitic protozoan Leishmania major. These repetitive elements share a 79-bp signature with other trypanosomatid non-LTR retroposons but are unique in that they are all

inactive, much more abundant, substantially conserved, and are located mainly in the 3'UTR of mRNAs. Conserved regions in 3'UTRs corresponding to SIDER1 elements have been reported to increase translation efficiency in a stage-specific manner. The SIDER2 subgroup has also been shown to regulate gene expression by promoting mRNA destabilization. Such observations prompted an in-depth comparative genomic analysis of SIDER and SIDER-like sequences. We report an optimal alignment of all SIDER sequences through probabilistic meta-algorithms for global optimization. The deterministic features of the poorly characterized SIDER1 were evaluated via phylogenetic analysis. A refined search profile for SIDER1 and SIDER2 was subsequently elaborated using Hidden Markov Models in order to achieve maximal discrimination during genomic scanning. The genomic distribution of search hits reveals that highly fragmented SIDERs are found throughout the genome and that fragments corresponding to particular SIDER positions are more abundant. Comparing SIDER1 and SIDER2 search results between L. major, L. infantum, and L. braziliensis reveal insightful clues on how these parasites may have assimilated transposable elements to gain an evolutionary edge.

OM7 : Wafae El Alaoui (UdeM)

DATATION MOLÉCULAIRE ET ARTEFACT DÛ À LA SATURATION MUTATIONNELLE

La phylogénie moléculaire fournit un outil complémentaire aux études paléontologiques et géologiques en permettant la construction phylogénétique des relations entre espèces ainsi que l'estimation du temps de leur divergence. Une calibration fossile est rajoutée à la phylogénie ce qui permet de déduire l'âge des autres nœuds de l'arbre, son principe repose sur l'hypothèse que les mutations apparaissent à intervalles réguliers. La théorie prédit plusieurs sources de variation dans les taux d'évolution moléculaire. Malgré cela, même une horloge approximative permet l'estimation du temps dans les différents événements de l'histoire évolutive, ce qui fournit une méthode pour tester plusieurs hypothèses sur l'origine de certaines espèces. Cependant lorsqu'un arbre phylogénétique est inféré, les chercheurs focalisent surtout sur la topologie, c'est-à-dire l'ordre de branchement relatif des différents noeuds. Les longueurs des branches de cette phylogénie sont considérées souvent comme des sous-produits, tandis que les longueurs des branches sont fonction de 2 paramètres : le taux de substitutions, et le temps. Les biologistes intéressés a étudier les taux évolutionnaires des génomes et taxa ou ceux qui veulent mesurer les temps de divergences entre espèces ont besoin d'une bonne estimation des longueurs de branches. On va donc s'intéresser à l'artefact du a la saturation mutationnelle et son influence sur la datation moléculaire. A ce jour, nous sous-estimons encore les longueurs des branches, surtout les longueurs relatives parce que nous ne détectons pas de façon efficace les substitutions multiples. Plus les séquences évoluent vite, plus grande est la sous estimation des substitutions multiples, plus forte est la saturation (c'est à dire la différence entre le nombre de substitutions actuelles et celles inférées). Ceci a pour conséquence de sous estimer la distance de la racine aux feuilles ainsi que son hétérogénéité. Quelques solutions s'offrent à nous pour améliorer la détection des substitutions multiples dont l'augmentation du nombre d'espèces afin de briser les plus longues branches de l'arbre, l'amélioration du modèle d'évolution des séquences afin de mieux décrire l'évolution des sites, ainsi que le retrait de sites à évolution rapide ce qui permet prendre en compte les imperfections des modèles actuels, et de ne garder que la fraction des sites qui vont le mieux s'ajuster au modèle spécifié. Ce projet a pour but de montrer l'impact de ces trois méthodes indépendantes sur l'estimation des longueurs de branches. Un jeu de données mammifères est utilisé pour l'ensemble de l'analyse puisqu'il possède un registre fossile riche et une saturation mutationnelle le long des branches très significative.

Oral Ph.D.

OD1 : Béatrice Roure (UdeM)

THE PROBABILITY OF IDENTICAL SUBSTITUTION PROFILES AS A CRITERION TO DETECT POSITIONS INVOLVED IN A FUNCTIONAL SHIFT

Since duplication is an important process that creates new function(1), the basic idea is to relax the criterion "constant but different"(2) by looking at amino acid positions that have a different substitional process in two paralogs. To determine these qualitative changes, a site-specific substitution profile is established by the CAT model(3), assuming that a gene evolves via different replacement processes. Two datasets (a and b haemoglobins, and the 7 proteins of the a proteasome) were used in the analysis and divided in three major monophyletic taxons. For each taxon, the substitution profiles were established and the probability that a site has the same profile in both taxa (called Probability of Identical Profile or PIP) was evaluated. The same protocol was applied to perform a parametric bootstrap test on simulated datasets. The number of positions with two different substitution profiles in orthologous comparisons is higher than expected by computer simlulations. This shows the existence of a qualitative heterogeneity across time. More importantly, the number of positions with two different substitution profiles is significantly higher between paralogs than between orthologs. Ideally, to determine sites involved in a functional shift, we looked for positions showing two different profiles between paralogs (PIP=0) and no difference among orthologs (PIP=1). The distributions of PIP values between ortho- and paralogous proteins are compared for each position, but only the a proteasome proteins show a large number of sites with a greater tendency to have two different profiles in pairwise paralogous comparisons. Nevertheless, the use of the PIP criterion is promising and we must improve the comprehension of these results by refining our detection strategy by combining them with other evolutionary criteria. (*The talk will be done in French)*

[1] Ohno S., 1970. Evolution by gene duplication. Springer-Verlag,, Berlin-Heidelberg-New York

[2] Gribaldo S. et al., 2003. Functional divergence prediction from evolutionary analysis: a case study of Vertebrate Hemoglobin. Mol. Biol. Evol. 20:1754-1759

[3] Lartillot N. and H. Philippe, 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 6:1095-109

OD2 : Mathieu Lajoie (UdeM)

EVOLUTION OF TANDEMLY ARRAYED GENES IN MULTIPLE SPECIES

Tandemly arrayed genes (TAG) constitute a large fraction of most genomes and play important roles. They evolve through unequal recombination, which places duplicated genes next to the original ones (tandem duplications). Many algorithms have been proposed to infer a tandem duplication history for a TAG cluster in a single species. However, the presence of different transcriptional orientations in most TAG clusters highlight the fact that processes such as inversion also contribute to their evolution. To circumvent this limitation, we proposed an extended evolutionary model which includes inversions and presented a branch-and-bound algorithm to infer a most parsimonious scenario of evolution for a given TAG cluster. Here, we generalize this model to multiple species and present a general framework to infer ancestral gene orders that minimize the number of inversions in the whole evolutionary history. An application on a pair of human-rat TAGs clusters is presented.

OD3 : Huiling Xiong (UOttawa)

USING GENERALIZED PROCRUSTES ANALYSIS (GPA) FOR NORMALIZATION OF CDNA MICROARRAY DATA

Normalization is essential in dual-labelled microarray data analysis to remove non-biological variations and systematic biases. Many normalization methods have been used to remove such biases within slides (Global, Lowess) and across slides (Scale, Quantile and VSN). However, all these popular approaches have critical assumptions about data distribution, which is often not valid in practice. In this study, we developed a novel assumption-free normalization method based on the Generalized Procrustes Analysis (GPA) algorithm. Both publicly available data and simulated data were used in comparing our GPA-based method with six other popular

normalization methods, including Global, Lowess, Scale, Quantile, VSN, and one boutique array-specific housekeeping gene method. The assessment of these methods is based on three different empirical criteria: across-slide variability, the Kolmogorov-Smirnov (K-S) statistic and the mean square error (MSE). Compared with other methods, the GPA-based method performs effectively and consistently better in reducing across-slide variability and removing systematic bias. In particular, the GPA method is free from the statistical and biological assumptions inherent in the other normalization methods that are often difficult to validate. The GPA-based method is therefore appropriate for diverse types of array sets, including the boutique array where the majority of genes may be differentially expressed.

OD4 : Yaoqing Shen (UdeM)

"UNITE AND CONQUER": ENHANCED PREDICTION OF PROTEIN SUBCELLULAR LOCALIZATION BY INTEGRATING MULTIPLE SPECIALIZED TOOLS

Knowing the subcellular location of proteins provides clues to their function as well as the interconnectivity of biological processes. Dozens of tools are available for predicting protein location in the eukaryotic cell. Each tool performs well on certain data sets, but their predictions often disagree for a given protein, which makes it difficult for the users to choose the right answer. Since the individual tools each have particular strengths, we set out to integrate them in a way that optimally exploits their potential. The method presented here is applicable to various subcellular locations, but tailored for predicting whether or not a protein is localized in mitochondria. Knowledge of the mitochondrial proteome is relevant to understanding the role of this organelle in global cellular processes. In order to develop a method for enhanced prediction of subcellular localization, we integrated the outputs of available localization prediction tools by several strategies, and tested the performance of each strategy with known mitochondrial proteins. The accuracy obtained (up to 92%) surpasses by far the individual tools. The method of integration proved crucial to the performance. For the prediction of mitochondrion-located proteins, integration via a two-layer decision tree clearly outperforms simpler methods, as it allows emphasis of biologically relevant features such as the mitochondrial targeting peptide and transmembrane domains. Our approach also alleviates the conundrum of how to choose between conflicting predictions. This approach is easy to implement, and applicable to predicting subcellular locations other than mitochondria, as well as other biological features. For a trial of our approach, we provide a webservice for mitochondrial protein prediction (named YimLOC), which can be accessed through the AnaBench suite at http://anabench.bcm.umontreal.ca/anabench/.

OD5 : Dapeng Zhang (UOttawa)

FUNCTIONAL INSIGHT INTO MAELSTROM PROTEINS IN THE GERMLINE SMALL RNA PATHWAY: A NOVEL DOMAIN WITH A DERIVED DNAQ 3'-5' EXONUCLEASE FOLD AND ITS LINEAGE-SPECIFIC EVOLUTIONARY EXPANSION/LOSS

Dapeng Zhang, Huiling Xiong, Xuhua Xia, Vance L. Trudeau

Germ plasm is considered as a cytoplasmic determinant for germline cell development. Many germline specific proteins have been identified to localize to this organelle and be functional in small RNA pathways including microRNA and recently discovered Piwi-interacting RNAs (piRNAs) or repeat-associated small interfering RNAs (rasiRNAs) pathways. Maelstrom protein is one key germ plasm-specific protein and function-related with Vasa and grk, however its function still remains uncovered. Here we tried to understand putative function of Maelstrom through using a combined strategy including sensitive sequence searching, evolutionary analysis, protein fold-recognition methods, and structural modeling. We found that a conserved functional module is exising in mael C-terminal part. Further fold recognition study showed that Mael-C domain bears a derived DnaQ exonuclease fold, which implies a nucleotide binding activity in RNAi pathway. Also evolutionary distribution showed this Mael-C domain exhibits lineage-specific expansions in several species, however was lost in all examined fish species. This in silico study not only provides a theory clue for maelstrom function in germ plasm and RNAi pathways of germline cells, but also demonstrates a clear protein evolution manner in which Mael-C domain bursts from the its ancient DnaQ domain through gene duplication and obtain new function through mutation in sequence whereas conservation in structural fold.

OD6 : Sivakumar Kannan (UMontréal)

UNASSIGNED MURF1 OF KINETOPLASTIDS CODES FOR NAD2

In a previous study, we conducted a large-scale similarity-free function prediction of mitochondrion-encoded hypothetical proteins, by which the hypothetical gene murf1 (maxicircle unidentified reading frame 1) was assigned as nad2, encoding subunit 2 of NADH dehydrogenase (Complex I of the respiratory chain). This hypothetical gene occurs in the mitochondrial genome of kinetoplastids, a group of unicellular eukaryotes including the causative agents of African sleeping sickness and leishmaniasis. In the present study, we test this assignment by using bioinformatics methods that are highly sensitive in identifying remote homologs and confront the prediction with available biological knowledge. Comparison of MURF1 profile Hidden Markov Model (HMM) against function-known profile HMMs in Pfam, Panther and TIGR shows that MURF1 is a Complex I protein, but without specifying the exact subunit. Therefore, we constructed profile HMMs for each individual subunit, using all available sequences clustered at various identity thresholds. HMM-HMM comparison of these of individual NADH subunits against MURF1 clearly identifies this hypothetical protein as NAD2. Further, we collected the relevant experimental information about kinetoplastids, which provides additional evidence for the in silico assignment of MURF1 being a highly divergent member of NAD2.

OD7 : Alireza Shaneh (McGill)

ON ACID-BASE COMPLEMENTARITY IN PROTEIN INTERACTION NETWORKS

Protein-protein interactions play a key role in many cellular processes. Protein interactions have been studied by their ability to be maintained (transient or permanent), their binding specificity, the regions implicated in interactions, and the similarity between interacting subunits. Therefore, it is essential to identify the physico-chemical properties which can contribute to protein nteraction networks. Acidic or basic activity of a protein represents one aspect of chemical properties in proteins. Consequently, can acidic and basic activities of proteins be considered as consistent indicators of protein-protein interactions? We studied the potential role of acid-base complementarity in protein interaction networks. We selected 18 experimentally shown etworks from Database of Interacting Proteins. To measure acidic or basic activity for each protein in the selected networks, we plotted protein charges as a function of pH using Henderson-Hasselbach equation. Our current results show that 44% of the selected networks follow acid-base complementarity concept. The acidic proteins in the selected networks interact with their partners which show basic activity. The applicability of acid-base complementarity in protein-protein interactions needs further investigation. The implication, interpretation and justification of our current results constitute the topic of this presentation.

OD8 : Karine Cyrenne (UdeM)

MODELISATION OF RNAS TRIDIMENSIONAL STRUCTURES (présentation en anglais)

Pendant les dernières années, les avancées en bioinformatique nous donnent accès à beaucoup d'information sur les différents types d'ARN et leurs rôles. Les ARNnc, entre autre, sont impliqués dans la régulation des gènes et maintes autres fonctions cellulaires. Les techniques de cristallographie à rayon X et résonance magnétique nucléaire permettant de déterminer la structure tridimensionnelle de l'ARN et aide à l'analyse de l'ARNnc. De plus, dans la cellule, l'ARN crée des interactions avec d'autres molécules. La structure tridimensionnelle permet de connaître les surfaces externes, donc accessibles, d'une séquence repliée ainsi que les surfaces disponibles pour l'interaction de la structure avec les autres molécules. Parallèlement aux méthodes de laboratoire, nous voulons automatiser une méthode qui prend en entrée une structure secondaire et donne comme résultats un ensemble de structures 3D correspondantes. La structure d'entrée fournie comme information la séquences des nucléotides et les interactions entre eux. Nous voulons fidèlement construire une structure ayant les relations spécifiées, avec des relations réelles, provenant de structures déterminées. Nous commençons d'abord par la construction d'une base de données. À partir des structures tridimensionnelles de la « Protein Data Bank », nous retenons les structures cristallographiées, ayant une résolution de 3 Angstrom ou moins. En date de novembre 2006, 519 structures ont été retenues. Pour chacune de ces structures, l'ensemble des relations, soient les nucléotides adjacents (linked), les paires canoniques et non-canoniques (paired), ainsi que les nucléotides

superposés (stacked) est extrait. Les structures sont modélisées à partir de cette base de données. L'implémentation de l'algorithme s'inspire de l'algorithme EM (« Expectation-Maximization »). Initialement, nous construisons une première structure aléatoire. Les relations choisies proviennent de différentes structures, alors les relations ne forment pas une structure parfaite. Il y a une erreur de construction que nous mesurons avec notre métrique de distance. Ensuite, nous améliorons cette première structure aléatoire en remplaçant les successivement les relations par des meilleures. L'erreur de construction est diminuée ou reste fixe. La boucle de maximisation se termine lorsque, pour une itération entière, toutes les relations de la structure restent inchangées. Connaissant l'erreur, nous avons une étape de distribution d'erreur, afin que chaque relation de la structure partage une fraction de l'erreur. Pour valider les performances de l'algorithme, nous avons construis des motifs et structures connues. Les premiers résultats sont des cycles de 4 nucléotides, soient la tetraloop GAAA et un tandem de pair de bases (GC/GA). Ceux-ci nécessitent en moyenne 2 à 3 itérations pour obtenir une structure optimale (5 sec. P6/1.6GHz). Les structures modélisées étaient près d'un modèle de référence (PDB :1AJF), avec des distances de 0.9 à 1.3 Angstrom. Nous avons ensuite considéré les deux cycles simultanément, soit une tige boucle minimale (2 paires de bases et la boucle GNRA). Encore une fois, l'algorithme converge, mais cette fois-ci en 5 ou 6 itérations (25 sec. P6/1.6GHz). L'algorithme que nous proposons converge rapidement et construit des structures plausibles. Nous pouvons ainsi générer un échantillonnage de conformations possibles, dont les relations proviennent de structures connues.

OD9 : Mathieu Coinçon (UdeM)

AU DELÀ DE LA MODÉLISATION: ÉTUDE DYNAMIQUE DE LA CATALYSE CHEZ LES FRUCTOSE-1,6-BISPHOSPHATE ALDOLASES

Obtenir la structure tridimensionnelle d'une enzyme n'est pas une fin en soit. Les données résultant de son étude ne sont généralement pas suffisantes à la compréhension de son mécanisme catalytique ou des interactions qui y sont liées. La flexibilité des structures protéiques et leur dynamique nous confrontent aux limites de nos méthodes d'études. Ainsi peu de méthodes expérimentales permettent d'étudier des phénomènes se déroulant dans des périodes de temps de l'ordre de la nanoseconde. Notre laboratoire se tourne alors vers des outils bioinformatique telle que la modélisation par homologie, l'arrimage moléculaire et les simulations de dynamique moléculaire pour compléter l'étude des structures cristallines obtenues. Les FBP aldolases de classe I et II nous offre une formidable plateforme pour tester la validité de ces méthodes et voir ce qu'elles peuvent nous apporter. Ces enzymes glycolytiques catalysent la réaction réversible de transformation du fructose-1,6- bisphosphate en glyceraldehyde-3-phosphate et en dihydroxyacetone-phosphate. Retrouvées principalement chez les organismes supérieurs pour les classes I et chez les inférieurs pour les classes II, leur étude pourrait conduire au développement d'antimicrobiens. Pour cela, leur mécanisme catalytique et les modalités de liaisons de leurs inhibiteurs spécifiques doivent être compris. Nous présentons ici la résolution de diverses structures tridimensionnelles d'aldolases obtenues par cristallographie et diffraction aux rayons X. Ainsi que l'étude d'inhibiteurs spécifiques de ses enzymes chez plusieurs organismes, ce qui nous a permis de valider des protocoles de simulation de dynamique moléculaire et d'arrimage moléculaire tout en explicitant certaines données expérimentales. Enfin, la simulation par dynamique moléculaire de boucles à rôle catalytique (non déterminées par cristallographie) nous a permis d'étudier et de comprendre le rôle de certains résidus mais aussi de mettre au point de nouvelles stratégies d'inhibition. Grâce à ces études, nous pouvons maintenant utiliser la modélisation à la fois comme illustration des données expérimentales mais aussi comme outils prédictif.

OD10 : Lilianne Dupuis (UdeM)

MULTISCALE SIMULATIONS OF PROTEIN FLEXIBILITY

Flexibility is an important property, inherent to the operation of a protein. Flexibility determines the way a protein interacts with other proteins or molecules. Static comparison methods between protein and ligand conformations do not take into account the mutual adaptation of their form when they get into contact. Molecular dynamic simulations may achieve a more realistic study, but computing time grows considerably when we increase the dimension of the molecules being simulated. The principal goal of our project is to combine speed and realism by using a multiscale approach, which simplifies the representation of the protein. A protein is organized in several levels. Its main chain folds into regular patterns: mainly α helices and β sheets. They

constitute the secondary structure of the protein and they alternate with irregular loops, more flexible. In the current project we consider the ordered and more rigid regions as single units, creating a higher level of approximation. An α helix or a β sheet is therefore considered as a single block, which may perform unified moves such as translations, rotations, swiveling, but also elastic deformations such as compressions, expansions or torsions. This allows us to simplify the computation of their motions and reserve computer time for the irregular regions, which perform more complex movements. However, in our implementation, computing time saving comes mainly from the fact that a much reduced number of possible conformation changes has to be taken into account in our seek for the molecular transformations. We use the activated method ART nouveau to find energetically favorable passages from one molecule configuration to the other. This technique is used in conjunction with the OPEP force field reformulated by our multiscale approach. The distribution between regular and irregular areas can be reevaluated on the fly after each event. Because realism is one of our main goals, we use a high difficulty test to evaluate our method. We start from conformations that are somewhat away from the native, and we observe how they come back to it. Several strategies have been studied and several development steps have been achieved toward an efficient operation.

Oral Post-Doctorat

OPD1 : Christian R. Landry (UMontréal)

AN IN VIVO MAP OF THE YEAST INTERACTOME

The systems-level behavior of cells is largely regulated through networks of protein-protein interactions. Deciphering the structure and dynamics of this network is therefore a central goal in biology and is a grand challenge for the field of functional genomics and bioinformatics. We have performed the first genome-wide in vivo and in intact cells screen of protein-protein interactions in Saccharomyces cerevisiae, a prime eukaryotic model. Our data is derived from an experiment consisting of more than 15 millions genetic crosses among strains tagged with reporters that allow the detection of binary and near-binary protein-protein interactions. We report with high confidence more than 2500 interactions. Integration of this in vivo network with previous large-scale experiments confirms several known interactions but most of the interactions we report are novel, due to the novel technique used (Protein-fragment complementation assay, PCA) and the nature of the assay, which allows us to cover a sub-space of the yeast interactome that has not been covered before. For instance, we discovered several novel protein-protein interactions among the proteins involved in autophagy, which is central to processes such as development and several human diseases. Our in vivo map of the yeast interactome will help better understand the physical behavior of proteins in their endogenous environment and how it maps to higher levels of organizations.

OPD2 : Véronique Ladret (UdeM)

STATISTICAL PROPERTIES OF ALLELIC HAPLOTYPE CLASSES AND STATISTICAL TESTS OF NATURAL SELECTION

The aim of my research project is to develop new statistical tests of neutral evolution in order to detect evidence of natural selection among DNA sequences. Several neutrality tests based on polymorphism data use the frequency distribution of haplotypes in the sample or site-by-site frequency spectrum. Our approach combines the polymorphism information by the means of allelic haplotype classes which classify the sequences in a given sample according to the number of sites at which they differ from the most recent common ancestor (MRCA) of the whole sample. In order to perform such tests, we first focus on the statistical properties of these classes. We assume a selectively neutral Wright-Fisher model of constant population size with an infinitely-many-sites model of mutation. For each i _ 0, we denote by Ci the class of sequences in a sample of size n that contain exactly i new mutations with respect to the MRCA. Let Ni be the number of haplotypes in class Ci. Using the coalescent, we derive the expectations, variances and covariances of the sizes of the classes Ci, i _ 0. These results allow us to calculate the theoretical mean and variance of the Sv statistic, formely introduced by Freytag et al. to detect signatures of natural selection among the sequence data, defined as some linear combination of the random variables Ni, i _ 0. The coalescent-based methods developed here might also apply to the calculation of the probability distribution of a given configuration which could lead to the development of new statistical tests of neutrality.

AFFICHES / POSTERS

Affiches M.Sc.

AM1 : Philippe Nadeau (UdeM)

ÉTUDE PAR SIMULATION DES SIGNATURES GÉNOMIQUES DE LA SÉLECTION NATURELLE

La variabilité du génome humain reflète l'ensemble de la diversité génétique retrouvée au sein de différentes populations. Cette variabilité est produite par des mutations créant des sites polymorphes représentés au niveau des chromosomes par différentes formes alléliques. La recombinaison, en redistribuant les allèles des sites polymorphes adjacents, crée de nouvelles combinaisons enrichissant ainsi la liste des haplotypes. Les fréquences alléliques et haplotypiques varient d'une population à l'autre et au cours des générations. En absence de la sélection naturelle, ces variations sont dues à la dérive génétique et aux évènements démographiques (croissance, migration, etc.). La sélection naturelle modifie la diversité génétique en changeant la distribution des fréquences alléliques des sites polymorphes ou des fréquences haplotypiques. C'est en comparant la diversité observée avec celle attendue sous neutralité que l'on retrouve des signatures génomiques de sélection naturelle qui peuvent être détectées à partir des spectres de fréquences alléliques ou haplotypiques modifiés. Ces changements de fréquences dépendent de la nature de la sélection. Différents indices mathématiques (statistiques descriptives) de diversité sont donc utilisés pour les retracer. Nous proposons d'investiguer une nouvelle statistique, appelée Sv2, comme outil d'identification des empreintes génomiques de la sélection naturelle. Sv2 dérive de la distribution des classes alléliques des haplotypes qui regroupent les haplotypes partageant le même nombre de sites polymorphes dérivés (non ancestraux). La statistique Sv2 devrait être sensible, à la fois, aux changements des fréquences haplotypiques et alléliques, se distinguant ainsi des autres statistiques utilisées dans les tests de neutralité et basées uniquement sur l'un ou sur l'autre. Pour étudier les propriétés de la statistique Sv2, je vais utiliser des données, sur la diversité génétique des populations, générées in silico par des programmes de simulations utilisant la théorie de la coalescence. Dans un premier temps, j'ai produit à l'aide du programme « ms » de Hudson des données représentatives du modèle d'évolution neutre. Grâce à ces données, j'explore différentes méthodes pour parcourir l'échantillon de séquences afin de calculer Sv2. Je compare aussi les valeurs de Sv2 pour les données sous neutralité avec celles calculées pour des données empiriques provenant de bases de données publiques. En second lieu, je devrai, plus tard au cours de mon projet, comparer la distribution des valeurs de Sv2 entre le modèle neutre et d'autres jeux de données. Le programme « cosi » de Schaffner permettra de simuler des données pour différents taux de mutation, taux de recombinaison et évènements démographiques. Alors que le programme « selsim » de Spencer et Coop sera utilisé pour générer des données soumises à la sélection naturelle. En bref, ce projet devrait déterminer l'impact particulier de la sélection naturelle sur la distribution de la statistique Sv2 à l'aide de simulations par ordinateur. Ce projet est en partie réalisé grâce au programme de bourses d'excellence biT.

AM2 : Martin Smith (UdeM / ULaval)

EVOLUTION, CHARACTERIZATION AND GENOMIC DISTRIBUTION OF TWO GROUPS OF SIDER IN LEISHMANIA PROTISTS

A large family of extinct retroposons known as SIDERs (Short Interspersed Degenerated Retroposons) has recently been identified in the genome of the parasitic protozoan Leishmania major. These repetitive elements share a 79-bp signature with other trypanosomatid non-LTR retroposons but are unique in that they are all inactive, much more abundant, substantially conserved, and are located mainly in the 3'UTR of mRNAs. Conserved regions in 3'UTRs corresponding to SIDER1 elements have been reported to increase translation efficiency in a stage-specific manner. The SIDER2 subgroup has also been shown to regulate gene expression by promoting mRNA destabilization. Such observations prompted an in-depth comparative genomic analysis of SIDER and SIDER-like sequences. We report an optimal alignment of all SIDER sequences through probabilistic meta-algorithms for global optimization. The deterministic features of the poorly characterized SIDER1 were evaluated via phylogenetic analysis. A refined search profile for SIDER1 and SIDER2 was subsequently elaborated using Hidden Markov Models in order to achieve maximal discrimination during

genomic scanning. The genomic distribution of search hits reveals that highly fragmented SIDERs are found throughout the genome and that fragments corresponding to particular SIDER positions are more abundant. Comparing SIDER1 and SIDER2 search results between L. major, L. infantum, and L. braziliensis reveal insightful clues on how these parasites may have assimilated transposable elements to gain an evolutionary edge.

AM3 : Julie Hussin (UdeM)

RECOMBINAISON MÉIOTIQUE ET STRUCTURE DU GÉNOME HUMAIN

La recombinaison méiotique est le principal mécanisme de fragmentation de la liaison génétique et est à l'origine de la diversité haplotypique du génome, redistribuant des nouvelles mutations parmi les chromosomes homologues. Grâce aux outils bio-informatiques et à l'abondance des données sur les variations haplotypiques dans les populations humaines, il est possible d'étudier précisément la distribution du taux de recombinaison ρ au niveau des séquences génomiques. L'objectif de l'étude est de décrire cette distribution le long de chromosomes humains et de caractériser les différences entre les populations. Pour ce faire, notre premier but est de construire un modèle qui décrirait la distribution neutre des recombinaisons, puis d'étudier les déviations observées et leurs causes. Comme point de départ, nous cherchons à modéliser la distribution des hotspots, petits segments génomiques très riches en recombinaisons, dans le génome humain. On montrera que, sous l'hypothèse de neutralité, les hotspots sont distribués le long des séquences selon une loi exponentielle. Sur les données empiriques cependant, l'hypothèse de neutralité est rejetée et les données sont décrites comme la somme de deux lois exponentielles. Étant donné les différences structurales connues entre les chromosomes du génome humain (par exemple : la taille, la position du centromère, etc.), nous avons ensuite tester lequel des deux modèles s'applique à chaque chromosome indépendamment. Ceci a été fait sur les données haplotypiques génotypées par Perlegen et traitées par Myers et al. 2005, grâce au logiciel LDhat, un programme d'analyse de patrons de déséquilibre de liaison et de recombinaison, dans le contexte de la théorie de la coalescence. Bien que la plupart des données par chromosomes s'expliquent par deux lois exponentielles, une seule loi s'applique mieux pour les chromosomes 10, 19, 22 et 23. L'étude des recombinaisons au niveau génomique s'inscrit dans les démarches visant à cartographier le génome humain et à élucider les causes génétiques des maladies multifactorielles.

[1] Myers et al. A fine-scale map of recombination rates and hotspots across the human genome. Science 310, pp.321-324 (2005)

[2] Lunter et al. Genome-Wide Identification of Human Functional DNA Using a Neutral Indel Model. PLoS Computational Biology, Volume 2, issue 1 (2006)

[3] Kong A. et al. A high-resolution recombination map of the human genome. Nature Genetics, Volume 31, pp.241-247 (2002)

AM4 : Xiaoquan Yao (UOttawa)

FUNCTIONAL DIVERGENCE IN METHIONINE AMINOPEPTIDASE (MAP) AMONG HUMAN, YEAST AND E. COLI: A BIOINFORMATIC APPROACH

N-terminal methionine excision (NME), carried out by the MAP enzyme, is a ubiquitous and conserved biochemical process, affecting the function of more than 50% of proteins in both prokaryotes and eukaryotes. NME efficiency is known to depend on the second and third amino acids (designated P1' and P2' amino acids, respectively), but it is experimentally difficult and tedious to accurately characterize the dependence of NME on P1' and P2' amino acids, making it a very tough task to study functional divergence among species. Here we present a new index to characterize this dependence of NME on P1' and P2' amino acids and use it to probe the functional divergence of MAP activities among Escherichia coli, the yeast (Saccharomyces cerevisiae), and human (Homo sapiens) representing prokaryotes, lower eukaryotes and higher eukaryotes, respectively. We found that the amino acids at both P1' and P2' sites of nascent proteins have significant effect on NME efficiency and that the three species have very different MAP activities. Our result explains why NME efficiency for peptides with Gly at P1' is relatively low in human. Five small and uncharged amino acids (Ala, Ser, Pro, Gly, and Thr) at P1' result in increased NME efficiency in all three species, but E. coli proteins with Val at P1' and human and yeast proteins with Cys at P1 exhibit poor NME efficiency. Our index helps identify

subtle differences in MAP specificity among human, the yeast and E. coli. When Ala is at P1', Asp, Glu and Lys at P2' increase NME efficiency in all three species, whereas Tyr, Ser, Phe, Asn and Leu decrease NME efficiency in all three species; when Ser is at P1', Glu at P2' increases NME efficiency in all three species, but Phe at P2' decreases NME efficiency in all three species.

AM5 : Diala Abd Rabbo (UdeM)

IDENTIFICATION DES PROFILS D'EXPRESSION ASSOCIÉS À UNE MUTATION DE L'UN DES GÈNES BRCA1 ET BRCA2 DANS LES CELLULES NON-TUMORALES DE L'ÉPITHÉLIUM DE SURFACE DE L'OVAIRE

L'identification des gènes BRCA1/2 représente une avancée majeure dans la compréhension de la carcinogenèse héréditaire. Les femmes porteuses de mutations sur l'un de ces gènes ont un risque élevé de développer des cancers de l'ovaire. Ces derniers, le plus souvent diagnostiqués à un stade avancé, seront mortels. Il est donc urgent d'améliorer les outils de diagnostic en identifiant des gènes différentiellement exprimés à un stade précoce de la carcinogenèse héréditaire. Des données de la littérature suggèrent l'existence d'un profil d'expression associé à une mutation des gènes BRCA1/2 dans les cancers de l'ovaire, ainsi que l'existence d'un phénotype particulier associé à l'hétérozygotie de ces gènes dans les cellules non-tumorales de l'épithélium de surface de l'ovaire (NOSEs). Nous avons émis l'hypothèse qu'il existait un profil d'expression différentiel associé à la présence d'une mutation dans les cellules non-tumorales avant même que n'apparaisse la tumeur. L'objectif de notre étude consiste donc à identifier le transcriptome associé à une mutation des gènes BRCA1/2 dans les cellules NOSEs. Nous disposions des données d'expression générées avec la biopuce Affymétrix HuFL® sur 9 échantillons d'ARN extraits de cultures primaires de NOSEs: 4 provenant de donneuses non-porteuses de mutation et 5 provenant de porteuses de mutation sur l'un des gènes BRCA1/2 définissant les Classes I et II respectivement. Nous avons effectué une analyse différentielle supervisée avec LIMMA (Linear Model for Microarrays Data) afin d'identifier les gènes différentiellement exprimés entre les échantillons des classes I et II. Nous avons ensuite appliqué une méthode d'agrégation hiérarchique aux gènes candidats sélectionnés par LIMMA pour vérifier si ces gènes permettent de séparer nos échantillons en fonction de leur classe d'appartenance. Nous avons finalement étudié la distribution des profils d'expression de nos échantillons grâce à une analyse de type RDA (Redundancy Analysis). Une validation biologique d'un sous-groupe de gènes candidats ainsi que la mesure de l'expression des gènes BRCA1/2 avec des méthodes de quantification en temps réel a été effectuée sur un plus grand nombre d'échantillons de NOSEs. Nous avons mis en évidence un profil d'expression différentiel associé à la présence d'une mutation dans les cellules NOSEs. La méthode LIMMA nous a permis d'identifier un ensemble de gènes candidats différentiellement exprimés entre échantillons non-mutés et mutés. Finalement, cet ensemble de gènes a permis de reclasser chaque échantillon dans la classe prédéfinie avant l'analyse supervisée. La validation biologique a permis de confirmer les profils d'expression obtenus pour un sous-groupe de nos candidats. À l'aide de la RDA, nous avons mis en évidence que nos échantillons ségrégeaient en 3 classes distinctes. Nous envisageons d'identifier les régions chromosomiques dont la dérégulation est associée à une mutation de l'un des gènes BRCA.

AM6 : Sam Khalouei (UToronto)

TRANSLATION INITIATION IN HUMAN IMMUNODEFICIENCY VIRUS TYPE 1: ANALYSIS OF HIV-1 5'-UNTRANSLATED REGION

The human immunodeficiency virus type 1 (HIV-1) is dependent on the host transcription and translation machinery to complete its life cycle. Different hypotheses have been postulated regarding the mechanism of translation initiation in HIV-1. It has been shown that HIV-1 mRNAs undergo cap-dependent ribosomal scanning and viral gene expression is modulated through varying signal strengths of the Kozak consensus sequences. Recent reports have shown evidence of cap-independent translation initiation mechanisms in HIV-1, such as the direct binding of ribosome at internal ribosome entry sites (IRES), distant from the 5' cap. The cap-dependent ribosomal scanning hypothesis predicts that there should be a selective pressure against ATG usage in optimal context in the HIV-1 5'UTR to avoid their erroneous selection by the scanning ribosome which hampers the detection of the true downstream translation initiation codon. The IRES-dependent translation initiation hypothesis, on the other hand, predicts that there is no such selective pressure since the ribosome directly binds at an internal site without having to scan the sequence. We evaluated these two hypotheses based

on their prediction regarding selective pressure against ATG usage in optimal context in the HIV-1 5'UTR. Our results show that there is indeed a selective pressure against such ATG's, which supports the cap-dependent translation initiation and contradicts the IRES-dependent translation initiation hypothesis. In the second part of our study we used a bioinformatics approach, including position weight matrix and perceptron, to analyze a region consisting of the 50 nucleotides upstream of the translation initiation codon in 331 unique HIV-1 and 24850 unique human sequences in an attempt to find conserved sites specific to HIV-1. We found that sites -17 and -25 (relative to the translation initiation codon) are highly conserved. Furthermore, using perceptron, we were able to separate HIV-1 and human sequences based on these upstream 5'UTR 50-mers. The strong site conservation specific to HIV-1 5'UTR 50-mers points to the possibility of as yet unknown cap-independent translation initiation mechanisms in HIV-1. The clear separation of HIV-1 and human sequences based on these 50-mers also raises the promising possibility of designing novel antiviral drugs that specifically suppress translation initiation in HIV-1 with little effect on human translation.

AM7 : Marie Pier Scott-Boyer (UdeM)

COMPUTATIONAL ANNOTATION OF NON-CODING RNAS IN CANDIDA ALBICANS

Candida albicans is a fungal pathogen causing systemic candidiasis, an important cause of mortality in immunocompromised patients. The annotation of the genome was performed with the exception of non-coding RNAs (ncRNAs). ncRNAs are involved in several essential processes in eukaryote cells for example: tRNA, snoRNA and rRNA. Antifungal drug discovery is limited by the availability of suitable drug targets and identifying essential ncRNAs would increase the pool of possible targets. ncRNAs are difficult to identify from genome sequence alone because secondary structure rather than sequence is responsible for their function. We scanned C. albicans haploid genome with the tools RFAM/Infernal [1], QRNA [2], RNAz [3] and Dynalign [4]. We will present results concerning the families identified by the different methods and the threshold scores used to determined significant hits. We will finally discuss the multiple strategies to prioritize predictions for validation. Known ncRNAs have mostly been identified in higher eukaryotes. Thus, to have optimal drug targets against fungi, we need a method to predict ncRNAs without using homology to a related organism or searching for a defined structure. To solve this problem, we propose an approach that enumerates (from genomic sequence) all possible compact motifs involving 1 or 2 base pairs that could form within all possible secondary structures. We then evaluate the ability to discriminate between structured and non-structured RNA sequences by comparing the distributions of enumerated motifs. After analyzing the results, we observed that most of the discriminating signal seems to come from a difference in dinucleotide composition in ncRNAs. We are currently developing an approach to use this information combined with a machine learning system (HMM) to recognize patterns present in the upstream region of genes to improve the prediction of ncRNAs. Preliminary results will be presented.

[1] Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy S R and Bateman A. Nucleic Acids Res., 33:D121-D124 (2005)
[2] Rivas E and Eddy S R. BMC Bioinformatics, 2:8 (2001)
[3] Washietl S, Hofacker I L and Stadler P F. Proc. Natl. Acad. Sci. U.S.A., 102:2454-2459 (2005)
[4] Mathews D H and Turner D H. J. Mol. Biol., 317:191 (2002)

AD1 : Nicolas Rodrigue (UdeM)

A MECHANISTIC ACCOUNT OF AMINO ACID OF CODON PREFERENCES IN MODELS OF PROTEIN-CODING NUCLEOTIDE SEQUENCE EVOLUTION

In 1994, Muse & Gaut (MG) and Goldman & Yang (GY) proposed attractive strategies for modeling the evolution of protein coding genes in a phylogenetic context. Their models recognize the coding structure of the sequences under study by defining a Markovian substitution process with a state space consisting of the 61 sense codons (assuming the universal genetic code). However, certain differences in the specification of MG and GY-style codon models have remained largely unaddressed, specifically regarding alternative parameterizations of the target states of substitution. On the one hand, we argue that the MG formulations offer more readily interpretable descriptions of the evolutionary process, clearly separating mutational parameterization from parameterization acknowledging the coding nature of the data. On the other, the MG models must contend with a rich GY formulation, based on a 61-dimensional frequency vector (GY-F61), which has been proposed as a practical way of accounting for the overall effects inducing uneven codon preferences. Unfortunately, as with other GY-style formulations, the GY-F61 model does not lend itself to an obvious biological interpretation, in this case confounding an account of mutational characteristics with other features leading to uneven codon preferences. In this work, we compare a set of MG and GY-style formulations using Bayes factors. We explore several variants along both axes, and propose new models in the MG-style that allow for a flexible account of global amino acid or codon preferences, while maintaining an distinct parameterization of mutational characteristics. From our analysis of three real data sets, we find that the different MG and GY approaches have a considerable impact on overall model fit, and that our proposed MG formulations accounting for amino acid or codon preferences tend to outperform or match the top-performing GY-style model. Altogether, the framework presented here suggests a broad modeling project in the MG-style, stressing the importance of combining and contrasting available model formulations, and grounding developments in a sound probabilistic paradigm.

AD2 : Yan Zhou (UdeM)

HANDLING HETEROTACHY WITH MBL MODEL AND COVARION MODEL

The evolutionary rate at a given homologous position varies across time. When sufficiently pronounced, this phenomenon—called heterotachy, may produce artefactual phylogenetic reconstructions under the commonly used models of sequence evolution. These observations have motivated the development of models that explicitly recognize heterotachy, with research directions proposed along two main axes: 1) the covarion approach, where sites switch from variable to invariable states; and 2) the mixture of branch lengths (MBL) approach, where alignment patterns are assumed to arise from one of several sets of branch lengths, under a given phylogeny. Here, we report the first statistical comparisons contrasting the performance of covarion and MBL modeling strategies. We also analyzed three large datasets (nuclear proteins of animals, mitochondrial proteins of mammals, and plastid proteins of plants). We demonstrated, using these large datasets, that the covarion model is more efficient at handling heterotachy than the MBL model. This is probably due to the fact that the MBL model requires a serious increase in the number of parameters, as compared to two supplementary parameters of the covarion approach. Further improvements of both the mixture and the covarion approaches might be obtained by modeling heterogeneous behavior both along time and across sites.

AD3 : Sivakumar Kannan (UdeM)

EVALUATING ORF FUNCTION PREDICTIONS USING DOMAIN-SPECIFIC KNOWLEDGE

Mitochondrial genomes from diverse eukaryotes carry on average 5 to 20 hypothetical proteins (ORFs) whose function has remained elusive. At the time of writing, GenBank stores more than 2,500 mitochondrial ORFs available from over 243 organisms. Using a machine learning based function prediction method, the decision tree algorithm C4.5 (Quinlan 1993), we have assigned function to 2,549 mitochondrial ORFs. However,

evaluation of these function assignments is not trivial. Performance evaluation methods such as ten-fold cross validation or leave-one-out only evaluate the performance of the classifier on the known but not on the unknown data. Sequence similarity based evaluation (multiple sequence alignment, for example) can also not be applied in this case, because these ORFs do not have recognizable homologs or domains. Therefore, we have developed criteria based on domain-specific knowledge to evaluate the function predictions of mitochondrial ORFs. The 'solitary rule' states that a mitochondrial genome contains only a single gene for each function. Consequently, a function prediction for an ORF is unlikely if a gene of the same function already has been found in the mtDNA of question. The second criterion is the 'solidarity rule' that corroborates a prediction if the proteome of a closely related species contains a protein with the same predicted function. Finally, the 'completeness criterion' considers whether a genome has been completely sequenced or not, in order to determine whether absence of a gene from the dataset means true absence from the genome. Using the above criteria, we were able to rank, out of 2,549 predictions, 614 as highly trustable, 759 as likely, and 1,152 as wrong predictions. As long as biochemical techniques do not lend themselves to high-throughput validation of function prediction, in silico predictions are of great value for wet-lab biologists, especially if all available evaluation criteria are taken together to establish most likely working hypotheses.

AD4 : Hamed Shateri Najafabadi (McGill)

PREDICTION OF PROTEIN INTERACTION MAPS FOR TRYPANOSOMA BRUCEI AND TRYPANOSOMA CRUZI: ALL FOR ONE, ONE FOR ALL

Trypanosoma brucei causes sleeping sickness in human and Nagana in livestock, while its nearest cousin, Trypanosoma cruzi is the cause of Chagas disease. Current drugs lack efficacy and cause severe side effects, and no vaccines are available. Increasing the available knowledge of the biology of the parasites is vital for the development of new drugs. The progress of various genome projects provides the opportunity to enhance the efficiency of traditional molecular biology approaches. Use of computer-aided and fully automated sequence analysis tools allows novel feature discovery as well as directed hypothesis-driven experiments. In this work, taking advantage of the complete genomic sequences of T. brucei and T. cruzi, we have exploited several protein-protein interaction prediction methods, including gene-fusion detection and phylogenic profile comparison to predict the protein interaction networks of these two trypanosomatids. To do so, we have compared the complete genome of these two parasites with the predicted protein sequences of 617 organisms whose complete genomes are deciphered. We have also used a Bayesian network which, by taking into account the codon usage and amino acid usage of interacting proteins, reduces the rate of false positive predictions. This network is trained on a gold standard set of interacting (positive) and non-interacting (negative) protein pairs and is used to refine the results of gene-fusion detection and phylogenic profile comparison methods as well as to expand the results to proteins for which no significant homology could be found in other organisms. The results show that each of these organisms has conserved interaction networks as well as some species-specific interactions. Many of these interactions are consisted of proteins of unknown function, providing new insights into their probable role.

AD5 : Yaoqing Shen (UdeM)

IN SILICO IDENTIFICATION OF THE MITOCHONDRIAL PROTEOME REVEALS DUAL LOCALIZATION OF BETA OXIDATION

In silico prediction of proteins imported into mitochondria is a powerful means to unravel the makeup of the mitochondrial proteome, the function of individual proteins, and the biological processes that take place in mitochondria. More than a dozen computational tools have been developed to predict the subcellular location of proteins, but their predictions for a given sequence often disagree. In order to make highly reliable predictions for mitochondrial proteins, we developed a new decision-tree based classifier, which integrates 13 available tools for subcellular localization and transmembrane domain prediction. This new classifier, which outperforms the individual predictors significantly, was applied to the newly sequenced genome of the fungus Rhizopus oryzae (Zygomycota). Among a total of 17,467 proteins, 1,766 were predicted being imported into mitochondria. By comparing the predicted mitochondrial proteome of Rhizopus with the experimentally confirmed one of Saccharomyces cerevisiae, we discovered that Rhizopus mitochondria host beta oxidation, a pathway which is absent from yeast mitochondria. In plants and fungi, beta oxidation is considered to operate

exclusively in peroxisomes. Specifically, proteomics studies show that this pathway is absent from mitochondria of the yeasts S. cerevisiae, Yarrowia lipolytica, and Candida tropicalis (Hiltunen et al., 1992; Kurihara et al., 1992; Smith et al., 2000). Yet, recently a mitochondrial form of beta oxidation was reported in Aspergillus nidulans (Ascomycota) and Sporidiobolus pararoseus (Basidiomycota) (Maggio-Hall & Keller, 2004; Feron et al., 2005). Still for Zygomycota, the subcellular localization of beta oxidation is unknown. Among the Rhizopus proteins of predicted mitochondrial localization, we discovered all the enzymes that catalyze the four steps of beta oxidation. This finding strongly suggests that in this zygomycete, beta oxidation takes place in mitochondria. When scrutinizing all non-mitochondrial proteins, we also discovered the peroxisomal components of beta oxidation (for example, the acyl-CoA oxidase which is predicted a peroxisomal localization). Therefore, we infer that in contrast to yeast, Rhizopus has both a mitochondrial and a peroxisomal beta oxidation, a situation similar to mammals. This finding illustrates the power of in silico predictions for the understanding of cellular metabolic networks, especially for species that are not amenable for experimental studies.

AD6 : Claude Bherer (UdeM)

CONTRIBUTION GÉNÉTIQUE DIFFÉRENTIELLE DES FONDATEURS AUX POOLS GÉNIQUES RÉGIONAUX DU QUÉBEC

Au Québec, les connaissances historiques, la distribution des maladies mendéliennes, de même que de récentes études génétiques suggèrent une stratification régionale du pool génique de la population contemporaine. L'analyse de généalogies ascendantes nous offre l'opportunité d'étudier les liens unissant les membres d'une population, à travers des générations d'ancêtres, jusqu'aux premiers immigrants sur le territoire. Afin d'évaluer l'impact de ces fondateurs sur la répartition de la variation génétique dans la population contemporaine, nous avons analysé un corpus de 2221 généalogies couvrant l'ensemble du territoire québécois, provenant de la base de données généalogiques BALSAC-RETRO. Ces généalogies ont une profondeur moyenne de 9,3 générations et atteignent jusqu'à 17 générations. Parmi les 153 477 ancêtres distincts apparaissant dans les généalogies, environ 7500 fondateurs ont été identifiés. La présente étude définit huit regroupements régionaux selon des mesures d'homogénéité (consanguinité) et de partage d'ancêtres (apparentement). La distribution et l'occurrence des fondateurs dans ces ensembles régionaux sont mesurées selon le sexe, l'origine géographique et la période d'arrivée au Québec. Enfin, la contribution génétique différentielle des fondateurs aux populations régionales est estimée selon ces mêmes caractéristiques. Des différences notables sont observées au niveau régional. Cette variabilité interrégionale concerne le nombre de fondateurs, leur diversité et leur contribution génétique, ainsi que l'apport d'immigrants arrivés plus tardivement qui affecte la composition et la structure génétique de ces populations.

AD7 : Lilianne Dupuis (UdeM)

MULTISCALE SIMULATIONS OF PROTEIN FLEXIBILITY

Flexibility is an important property, inherent to the operation of a protein. Flexibility determines the way a protein interacts with other proteins or molecules. Static comparison methods between protein and ligand conformations do not take into account the mutual adaptation of their form when they get into contact. Molecular dynamic simulations may achieve a more realistic study, but computing time grows considerably when we increase the dimension of the molecules being simulated. The principal goal of our project is to combine speed and realism by using a multiscale approach, which simplifies the representation of the protein. A protein is organized in several levels. Its main chain folds into regular patterns: mainly α helices and β sheets. They constitute the secondary structure of the protein and they alternate with irregular loops, more flexible. In the current project we consider the ordered and more rigid regions as single units, creating a higher level of approximation. An α helix or a β sheet is therefore considered as a single block, which may perform unified moves such as translations, rotations, swiveling, but also elastic deformations such as compressions, expansions or torsions. This allows us to simplify the computation of their motions and reserve computer time for the irregular regions, which perform more complex movements. However, in our implementation, computing time saving comes mainly from the fact that a much reduced number of possible conformation changes has to be taken into account in our seek for the molecular transformations. We use the activated method ART nouveau to find energetically favorable passages from one molecule configuration to the other. This technique is used in

conjunction with the OPEP force field reformulated by our multiscale approach. The distribution between regular and irregular areas can be reevaluated on the fly after each event. Because realism is one of our main goals, we use a high difficulty test to evaluate our method. We start from conformations that are somewhat away from the native, and we observe how they come back to it. Several strategies have been studied and several development steps have been achieved toward an efficient operation.

AD8 : André Levasseur (UdeM)

INFÉRENCE D'ORTHOLOGIE PAR CONTEXTE GÉNOMIQUE

La qualité de l'assignation de gènes orthologues est un pré requis essentiel à plusieurs secteurs clés de la génomique comparative tels que la prédiction de fonctions, la construction d'arbres phylogénétiques d'espèces et l'analyse des réarrangements génomiques. La justesse d'une prédiction d'orthologie basée sur la comparaison réciproque de séquences (BBH) est affectée par la présence de pseudo-orthologues et de xénologues qui génèrent tous deux des faux positifs. Les pseudo-orthologues s'obtiennent par exemple lorsqu'un gène ancestral et sa copie sont alternativement perdus entre deux lignées, de sorte qu'en comparant les deux génomes actuels, on ne retrouve dans l'un qu'une copie et dans l'autre que le gène ancestral. Les xénologues s'obtiennent lorsqu'un orthologue phylogénétiquement distant est acquis par transfert horizontal de gène (THG) et que le gène ancestral est par la suite perdu. Notre travail repose sur l'hypothèse que les paires de gènes ayant conservé un voisinage génique similaire (homologues positionnels) sont plus susceptibles d'être des orthologues. Ainsi, dans les cas où le voisinage a suffisamment survécu aux différentes mutations à grande échelle, les paires d'homologues positionnels alors identifiées auront de fortes chances d'être de vrais orthologues, évitant le piège des pseudo-orthologues et des xénologues. Nous présentons un nouvel algorithme basé sur les intervalles communs qui, pour une paire de génomes donnée, permet d'identifier toutes les paires d'homologues positionnels. Cet algorithme, rapide et efficace, traite d'abord les cas évidents de conservation de voisinage à l'aide de segments communs et par la suite, utilise les intervalles communs pour les autres cas moins évidents. Nous présentons aussi les résultats obtenus avec des paires de génomes bactériens de distance phylogénétique croissante. Une version de cet algorithme capable de traiter un ensemble de génomes est en cours de développement. Nous tenterons par la suite de démontrer que les orthologues ainsi obtenus permettent de calculer de meilleures phylogénies.

Affiches hors concours

AHC1 : Elisabeth Tillier (UToronto)

A FAST AND FLEXIBLE APPROACH TO OLIGONUCLEOTIDE PROBE DESIGN FOR GENOMES AND GENE FAMILIES

Motivation: With hundreds of completely sequenced microbial genomes available, and advancements in DNA microarray technology, the detection of genes in microbial communities consisting of hundreds of thousands sequences may be possible. The existing strategies developed for DNA probe design, geared toward identifying specific sequences, are not suitable due to the lack of coverage, flexibility and efficiency necessary for applications in metagenomics. Results: ProDesign is very flexible in that it can be used for designing probes for detecting many genes families simultaneously and specifically in one or more genomes. We have found that ProDesign provides more flexibility, coverage and speed than other software programs used in the selection of probes for genomic and gene family arrays.

AHC2 : Alexandro Murua (UdeM)

MODELING REPLICATED TIME-COURSE EVOLUTION OF GENE-EXPRESSION PROFILES

Time course studies (based on repeated measures over time) allow investigating the dynamic evolution of complex processes. Such experimental designs have been widely used in many biological studies and are of particular interest in gene expression profiling experiments. For technical, experimental and biological reasons these studies produce rather noisy signals. In consequence it is common practice to use replicates to capture some of the uncertainty. Here we present a statistical model for time-course studies including (independent) replicates. Our model for analysis deals with smoothed time-courses through a special case of a Gaussian intrinsic autoregressive process. The smoothed time-courses are obtained by penalizing the size of their second derivates. The resulting model is embedded in a Bayesian framework. Parameters estimates are obtained via Iterated Conditional Modes (ICM). If replicates are available, we estimate the smoothing parameter (that penalizes the second derivatives) using m-fold cross-validation. This methodology is applied to the analysis and clustering of gene-expression data. Our opposed to working with the original signal plus noise time-courses) lead to better clustering results.

AHC3 : Emmet O'Brien (UdeM)

TBESTDB: A TAXONOMICALLY BROAD EST DATABASE

TBestDB (http://tbestdb.bcm.umontrealca) is a repository for EST data from a taxonomically broad range of mostly unicellular eukaryotes, many of which have not previously been thoroughly investigated. Much of this information was gathered by the Protist EST Program (PEP), a collaboration between six Canadians laboratories. TBestDB currently (April 2007) contains ~470,000 publicly available ESTs from 55 organisms, which allows both, studies focused on individual organisms, and exploration of fundamental biological questions across a taxonomically diverse dataset. We offer TBestDB as a service for clustering, annotation and distribution of EST data. Password-controlled access can be provided to data undergoing processing for a limited period of time.