Compte-rendu des journées NETBIO des 18/19 septembre 2014

Jeudi 18 septembre matin

10h30-11h30 : Richard Berthome *Rôle des petits ARNs dans la régulation de l'expression des gènes*

La présentation commence par une revue historique de la découverte du rôle des petits ARNs dans les années 90 : ARN tige boucle qui silence l'expression d'un gène, transport, amplification et action post-transcriptionnelle. Suit une présentation plus spécifique des siRNA (ARN double brin, pris en charge par des ANR polymérases qui amplifient leur signal) et des miRNA ou hpRNA (21/24 nucléotides, double brin avec tige boucle ; qui sont sélectionnés au cours de l'évolution et ciblent un transcrit pour iniber sa traduction). Les différents types d'action des petits ARNs chez les eucaryotes sont aussi présentés : dégradation d'un ARNm par appariement ou inibition lorsque l'appariement est imparfait (régulation post-transcriptionnelle) ou bien régulation transcriptionnelle par inibition de la traduction. La recherche des petits ARNs est faite par approche ARNomique ou in silico (méthodes de prédiction) et sont ensuite validés par diverses approches expérimentales (puces ADN, recherche d'ANRm cible...). Les cibles des petits ARNs sont diverses mais concernent la fonction de transcription principalement (études chez *Arabidopsis*). Un exemple est donné sur un miRNA qui régule le passage à la phase adulte (reproduction) chez la plante. Enfin, des techniques d'identification des petits ARN sont présentées (infection virale, expression d'un ARN tige-boucle artificiel...).

11h30-12h00 : Amine Ghozlane *Genome-wide detection of sRNA targets with rNAV / présentation de Tulip*

La présentation concerne une approche de prédiction des micro-RNA avec des méthodes de visualisation (utilisant le logiciel Tulip). L'étude porte sur les sRNA (bacterial small RNA) qui est le processus principal de régulation de l'expression chez les bactéries et qui interviennent particulièrement en condition de stress. Des centaines de sRNA ont été identifiés par séquençage par les cibles connues sont pour l'instant très limitées et la caractérisation de la fonction est un défit important (particulièrement sur un contexte haut débit). Il a été montré qu'un petit ARN intervient sur un ensemble de transcrits qui appartiennent à la même voie méthabolique : le but du travail est d'intégrer les méthodes de prédiction des interactions sRNA-mRNA, avec une caractérisation fonctionnelle et une méthode de visualisation de graphes bipartis (sRNA-mRNA).

12h00-12h20 : Oriol Guitart Pla *The Cytoscape Network Inference (Cyni) toolbox for gene regulatory network inference*

La présentation concerne la boîte à outil Cyni de Cytoscape pour l'inférence de réseau de régulation. Elle commence par une vision d'ensemble de Cytoscape. Cyni contient plusieurs méthodes d'inférence, de plusieurs méthodes d'imputation et de discrétisation, de plusieurs métriques... Il existe une documentation complète et des tutoriels. Cyni est disponible sur l' « App store » de Cytoscape (libre) et possède une interface utilisateur interactive. Plusieurs méthodes issues de la littérature sont implémentées, par exemple « ARACNE » (Califano *et al.*) qui est une approche LASSO permettant d'intégrer des a priori biologiques à partir d'un processus de diffusion qui a gagné un des challenges DREAM.

Jeudi 18 septembre après-midi

14h00-15h : Pierre Barbillon *Network impact on persistence in a finite population dynamic exchange model*

La problématique est celle de l'influence d'un réseau social (entre paysans) sur la préservation de

certaines variétés anciennes. Les données sont un graphe qui correspond aux échanges de graines entre les paysans qui est réputé être l'observation issue d'un réseau de conseil et d'échange non observé. Le modèle est un modèle dynamique probabiliste (proche du modèle SIR) où chaque sommet correspond à une ferme, qui contient ou non une variété donnée. Pour un nombre faible de fermes, les calculs exacts peuvent être effectués et pour un nombre de fermes plus élevés et à temps fini, une analyse de sensibilité sur des simulations est effectuée pour comprendre l'influence du réseau sur la probabilité d'extinction à temps fini. Différentes techniques de simulation sont présentées pour estimer les probabilités d'évènements rares (extinctions rares, persistances rares...). Différentes topologies de réseaux sont comparées (Erdos Rényi, SBM, lattice model, preferential attachment model) avec une ANOVA sur les divers paramètres des modèles de simulation.

15h00-15h30 : Rim Zaag *From gene expression modelling to coregulation networks for Arabidopsis*

La présentation est motivée par la notation fonctionnelle des gènes : même pour Arabidopsis, plus de 5000 gènes sont sans fonction identifiée. Une voie développée actuellement est de se baser sur la co-expression des gènes pour aider à l'anotation en effectuant une classification non supervisée. Le cadre d'étude est l'étude de données d'expression chez *Arabidopsis* dans des conditions de stress diverses. Les expressions gènes différentiellement exprimés sont traitées par une méthode de classification non supervisée basée sur un modèle statistique : cette méthode a été implémentée dans GEM2Net dans CATdb : elle associe les résultats de la classification avec des informations d'ontologie et permet l'inférence de réseaux pour chacune des classes de co-expression. Une approche par simulation est utilisée pour déterminer qu'une 4 co-expression dans au moins 7 conditions de stress est une bonne indication de co-régulation des gènes. Des analyses d'enrichissement fonctionnel et des analyses d'enrichissement en motifs Cis-régulateurs sont utilisées pour valider les résultats.

15h30-16h00 : Yann Vasseur *Régressions pénalisées et modèles à blocs latents pour une caractérisation relationnelle des facteurs de transcription d'Arabidopsis*

La présentation concerne l'étude des facteurs de transcription. Celle-ci est menée par le biais d'expériences transcriptomiques. Les liens de corrélation entre facteurs de transcription sont estimés dans le cadre d'un modèle graphique gaussien (corrélation partielle). Pour effectuer l'estimation, une approche dite « Gauss LASSO » est utilisée avec le choix du paramètre de régulation par BIC. Elle est combinée à une approche par ré-échantillonnage pour limiter le nombre de facteurs de transcription potentiels. L'étude des signes des coefficients du modèle de régression montre que la plupart des facteurs de transcription ont une action activatrice exclusivement.

16h30-17h30 : Actualités de NETBIO et discussions

Matthieu Vignes et Simon de Givry ont quitté l'animation du réseau et ont été remplacés par Julien Chiquet Nathalie Villa-Vialaneix.
Le financement provient aux 2/3 du département MIA et pour 1/3 du réseau national de systèmes complexes. Cette année, 9000 € ont été obtenus par ce biais. Le financement permet aussi de prendre en charge le déplacement des Toulousains. Le réseau national des systèmes complexes demandent à ce que les participants apparaissent dans leur annuaire : ce point est encore à faire.
Le site web (page web en fait) est maintenu par NV² et contient l'historique des journées. Le site est en voie de migration par JC vers un hébergement spécifique par le département MIA.
Patrick Meyer (UCL) va venir du 24 au 27 novembre en invité. Ceux qui sont intéressés par le rencontrer peuvent se faire connaître pour qu'un programme soit organisé autour de sa venue.
Une autre idée pour utiliser les fonds collectés pourrait être d'organiser des visites pour les doctorants entre unités participant à NETBIO.
Une autre question posée au réseau NETBIO est la demande de biologistes pour des formations sur l'inférence de réseaux : NETBIO a-t-elle vocation à s'intéresser à cette question ? Comment ? Le

constat est que certains biologistes se lancent dans l'inférence de réseaux et sont plus visibles que les modélisateurs. Sur cette question, il semblerait que la formation permanente soit en train de tenter de créer une formation sur les réseaux. Une autre manière de compiler la connaissance accumulée dans le réseau serait de répondre à la sollicitation d'un éditeur pour un livre.
Un sujet COST avait été posé par Matthieu sur la création de modèles de simulation pour tester des méthodes d'inférence de réseaux et il avait été rejeté par manque d'implication de biologistes. Une remarque est qu'il est peut-être nécessaire d'avoir une visibilité nationale plus forte. MLMM fait remarquer que avec les archives des journées NETBIO, il y a déjà pratiquement le matériel pour monter une formation : les connaissances accumulées suite aux nombreuses discussions lors des journées pourraient être mises à profit lors d'ateliers ou de formations (qu'est-ce qu'un réseau ? pourquoi ?...). NETBIO pourrait être pionner pour montrer des problèmes concrets où le modèle d'inférence a apporté.
Est-ce que la manifestation NETBIO pourrait devenir annuelle ? Une session introductive avec une présentation généraliste pourrait être prévue systématiquement. Cette session pourrait être une session par des biologistes ou une session par des statisticiens. La demande des biologistes aux statisticiens est d'avoir des sessions introductives à une méthodologie donnée, type « cours ».
Il n'y a pas de retour ni de projets montés par NETBIO mais il y a des discussions et des collaborations en cours ou suscitées par le réseau. AIGM organise un workshop avec des lectures d'articles et des présentations de travaux de doctorants.
Une deuxième session pourrait être organisée vers mars. Le réseau pourrait être étendu à un peu plus de biologistes en les sollicitant plus directement. Cette manifestation pourrait être plus restreinte sur le temps (1 jour) et organisée autour de présentations thématiques et/ou des questions ouvertes en biologie, statistique (type « workshop »).

Vendredi 19 septembre matin

9h00-10h00 : Pierre Latouche *Random graph models for the clustering of nodes in networks and visualisation*

La présentation se focalise sur des méthodes pour explorer / analyser des graphes déjà connus : classification, visualisation. Du point de vue de la classification, on distingue les méthodes de recherche de communautés (recherche de groupes denses) et les méthode « dissassortative mixing », ces deux approches étant généralisées dans la recherche de structures hétérogènes. Du point de vue de la visualisation, certaines des méthodes projettent le graphe dans un espace latent de faible dimension : des approches combines cette visualisation avec une classification. Une méthode fréquemment utilisée pour réaliser la classification de sommets est d'utiliser un modèle stochastique à blocs latents (« stochastic block model »). Une approche par EM variationnelle pour l'optimisation de la vraisemblance est décrite pour la classification de sommets dans ce modèle et des résultats sont montrés pour un réseau métabolique provenant de *E. Coli*. Une généralisation de ce modèle permet de trouver des classes chevauchantes. L'approche est illustrée sur des données simulées et des données issues d'un réseau de régulation de la levure. Elle permet de retrouver des groupes cohérents qui correspondent à des motifs de réponses à des stress. Enfin, les W-graphes (graphons) sont présentés : il s'agit de modèles de graphes aléatoires basées sur une structure de surface continue : une méthode d'estimation des paramètres du modèle à partir de données observées est proposée : elle est également basée sur le modèle stochastique à blocs latents.

10h00-11h00 : Fabrice Rossi *Triclustering pour la détection de structures temporelles dans les graphes*

La présentation concerne des données d'interactions temporelles : un acteur émet ou reçoit un message vers un autre avec une étiquette temporelle sur l'interaction (liste de sources et de récepteurs). Ce type

de données peut-être modélisé par un graphe avec une fonction de présence qui indique si une arête donnée du graphe est présente à l'instant *t* (fonction à valeur dans {0,1} et arête instantanée). Deux idées peuvent être mises en œuvre pour effectuer des classifications de sommets dans ce type de données : la première consiste à supposer que les schémas d'interactions sont fixées et que les appartenances aux classes varient. La deuxième, utilisée ici, fixe la partition des sommets et fait varier les schémas d'interaction entre les classes. Un modèle génératif est proposé pour générer un jeu de données simulé pour des paramètres du modèle (partition des sommets et schémas d'interactions) par une approche hiérarchique avec distribution uniforme à tous les niveaux hiérarchiques. Une approche EM est utilisée pour retrouver la partition des sources, des destinations et des temps pour produire des classes de sommets pour des « instantanés temporels » combinant un intervalle de temps. Le problème de l'optimisation du critère est un problème combinatoriel compliqué qui est résolu avec des heuristiques destinées à diminuer la complexité de la résolution (approche gloutonne avec raffinement multi-niveaux). L'approche est testée sur des données simulées et des données réelles sur les appels téléphoniques en Côte d'Ivoire.

11h30-12h00 : Trung Ha *Régressions multivariées en grande dimension*

Le contexte de cette présentation est l'inférence de réseau avec un modèle graphique gaussien (GGM) dans le contexte où les données proviennent de plusieurs échantillons. Le modèle proposé inclut trois types de pénalité, une qui côntrôle le nombre total d'arêtes (et/ou les différences entre conditions), une deuxième qui contrôle la différence entre les moyennes des expressions du gènes dans chaque condition et une dernière qui contrôle la magnétude de l'estimation de la moyenne de l'expression du gène dans chaque échantillon. L'approche est testée sur des données simulées, en évaluant à la fois la qualité de l'estimation des moyennes et de la reconstruction du réseau.

12h00-12h30 : Mélina Galopin *Model-based clustering of genes expression data with external annotations*

Cette présentation se place dans le contexte des données RNAseq. Ces données sont des données de comptage et les méthodes utilisées pour l'analyse de données de puces ne sont pas directement applicables (distribution non gaussienne). Les alternatives qui existent pour traiter ce type de données sont soit d'appliquer une transformation des données initiales pour appliquer les modèles gaussiens, soit de modéliser les données de comptage par des lois de Poisson ou des lois binomiales négatives et de construire un modèle adapté à cette hypothèse. La présentation se concentre sur la classification de gènes à partir des données d'expression et des données d'annotation GO : les données d'annotation sont utilisées dans l'étape de sélection de modèle d'un modèle de mélange pour la classification à partir des données d'expression. Un critère SICL est utilisé qui combine un critère ICL et un critère lié à une classification externe. La méthode est illustrée sur un jeu de données réel provenant d'une étude sur le porc.

12h00-12h30 : Loic Schwaller *Tree-based graphical model inference*

La présentation concerne l'inférence de réseaux quand on suppose une sous-structure en arbre. Loïc s'intéresse à l'inférence en tenant compte de cette information. En introduction, Loïc a présenté l'algorithme de Chow & Liu qui permet de trouver l'arbre recouvrant maximal par la sélection d'un arbre dans la collection d'arbre considéré. Dans son travail, Loïc propose une inférence qui prend en compte toute l'information des inférences obtenues sur la collection d'arbres considérées. Le résultat de l'inférence est la probabilité a posteriori de la présence d'une arête. Les 2 méthodes sont comparées sur une première étude de simulation en terme de VP et FP puis dans une seconde étude de simulation, l'estimation par calcul exact sont comparés à l'estimation par MCMC. Les calculs exacts sont plus rapide.

15h00-16h00 : Avner Bar-Hen *Influential observations in a Graphical Model*

Avner commence son exposé par un rappel sur un GGM, son interprétation et l'inférence. Pour illustrer la méthode, le jeu de données de Hess et al (2006) est utilisé et présenté au début de la présentation. La question est de connaître la stabilité de l'inférence en étudiant la fonction d'influence des observations. Pour fixer les notations et techniques utilisées, Avner fait également un rappel sur les fonctions d'influence et montre que l'on peut recalculer la matrice de variance-covariance des (n-1) observations à partir de la matrice de variance-covariance calcule avec les n observations. Dans l'exposé, Avner propose plusieurs index pour mesurer l'influence de chaque observation et discute les performances et leur pertinence sur le jeu de données de Hess et al (2006). Premier indice sur la matrice d'adjacence et le second indice sur la vraisemblance. Avner discute aussi leur approche par rapport à Bolasso.

16h30-17h00 : Jean-Michel Bécu *Sélection de variables par la ridge adaptative*

La présentation concerne le test de la significativité dans les régressions. Quand on est dans un cadre de la grande dimension, les tests usuels ne sont pas utilisables. Dans le cadre de la régressions ridge, Jean-Michel propose une procédure en 2 étapes. La première concerne le screening pour identifier les variables importantes pour la régression, la seconde étape concerne le cleaning. Elle permet d'identifier les variables dont le coefficient est significativement différent de 0. Cette approche est évaluée et comparée à des méthodes concurrentes sur un ensemble de données simulées.