Compte-rendu des journées NETBIO qui ont eu lieu le 19 et 20 novembre 2012 à l'Institut Pasteur

Simon de Givry, Matthieu Vignes et Marie-Laure Martin-Magniette

Les journées ont rassemblé une trentaine de personnes. Les exposés sont en ligne sur le site du réseau à l'adresse suivante :
http://carlit.toulouse.inra.fr/wikiz/index.php/Inférence_de_réseaux_-_réseau_MIA

Anne-Claire Haury a présenté TIGRESS dont l'objectif est d'identifier les cibles des facteurs de transcription à partir de données d'expression. Le principe est de faire une analyse gène par gène pour régresser l'expression du gène sur l'expression des facteurs de transcription. La méthode de régularisation utilisée est le LASSO (algorithme LARS) ; un bootstrap est réalisé pour stabiliser les résultats. Ceci permet de calculer une fréquence de sélection pour chaque facteur de transcription. Lors des challenges DREAM5, TIGRESS est arrivée troisième, derrière une méthode permettant d'avoir des interactions entre facteurs de transcription et une méthode utilisant des forêts aléatoires pour le choix des régresseurs dans le modèle linéaire.

Anne Siegel nous a expliqué la programmation par ensembles-réponses (ASP) sur un exemple simple de Cluedo. Le principe est d'énumérer toutes la connaissance dont on dispose sur un système puis d'utiliser un solveur pour énumérer l'ensemble des configurations qui sont solutions. Anne a ensuite présenté ses travaux autour de l'utilisation de la programmation par ensembles-réponses pour des questions autour de l'inférence de réseaux. Ceci permet de valider et/ou corriger des réseaux en cours de construction, d'en inférer ou de les compléter.

Simon de Givry nous a présenté les réseaux bayésiens statiques. Les noeuds sont les gènes qui sont décrits par des variables booléennes et l'objectif est de reconstruire le réseau qui lie les gènes. Les méthodes pour la reconstruction de réseau bayésien sont ici fondées sur une fonction objective à maximiser, appelée score. Il doit être facile à calculer et ne doit pas faire du sur-apprentissage. Les scores proposés sont fondés sur la vraisemblance du modèle est sont la somme de score locaux, ce qui permet une recherche itérative.
Simon a décrit les différentes étapes nécessaires pour la définition d'un algorithme et a présenté des améliorations proposées lors de la thèse de Jimmy Vandel sous forme de nouveau mouvements dans l'espace des opérateurs sur graphes.

Arnaud Fouchet a présenté ses travaux de thèse sur la modélisation d'un réseau ayant une dynamique non linéaire à l'aide de méthodes à noyaux. L'idée est de pondérer plusieurs noyaux unidimensionnels. Les données utilisées sont issues d'une cinétique. Pour évaluer la stabilité, un bootstrap en utilisant des sous-ensembles de données construits pour garder l'idée de la cinétique est proposé.

Mélina Gallopin a présenté les travaux de stage de master sur l'inférence de réseaux à partir de données RNAseq. Depuis 2008, les transcriptomes peuvent être observés grâce des technologies de séquençage haut-débit qui permettent d'avoir un comptage de transcrits au lieu d'un signal 'hybridation. Les données sont donc discrètes et positives et il est fréquemment observé que la variance empirique eest supérieure à la moyenne empirique. Lors de son stage, Mélina a comparé 3 méthodes d'inférence de réseau (i) un modèle graphique Gaussien après avoir effectué une transformation log(y+1) (ii) un modèle de Poisson log-lineaire, où les données sont les comptages élevés à une certaine puissance. Ce modèle a été proposé par Allen et Liu (2012) (iii) un modèle hiérarchique Poisson log-normale qui ne nécessite pas de transformation des données. Les modèles ont été comparés sur des données simulées à partir de lois de Poisson multivariées et le modèle hierarchique semble le meilleur.

Pierre Guttierrez a commencé son stage de master depuis 2 mois ½ et a présenté ses premiers résultats. L'objectif est d'identifier les acteurs d'un réseau et les relations entre les acteurs simultanément. Pour cela, il propose une procédure en 2 étapes qui consiste à faire une sélection des gènes à considérer et un modèle graphique Gaussien. Pierre nous a présenté la première étape, l'idée est de réécrire l'analyse différentielle comme une régression pénalisée en utilisant une fused ANOVA.

Pierre Nicolas a exposé les travaux du projet BasysBio où a été produit 269 hybridations sur la vie de Bacillus et 104 hybridations sur différentes conditions de stress avec 2 ou 3 réplicats biologiques. Les données ont été produites à l'aide d'une puce tiling avec une sonde qui débute tous les 22 nucléotides. Des méthodes de segmentation ont été utilisées pour trouver les régions transcrites avec un modèle HMM où la variable latente est continue mais discrétisée. L'objectif était de bien modéliser le signal pour ensuite l'interprétation. Pierre nous a aussi présenté la création d'un catalogue de transcrits à partir de promoteurs déjà connus et une classification des promoteurs fondée sur du clustering hiérarchique. Pour finir, Pierre a montré l'activité des anti-sens.

Françoise Monéger a expliqué la méthodologie utilisée pour comprendre le réseau génique qui contrôle le développement floral. Depuis longtemps, beaucoup d'informations disponibles et la question était de savoir s'il était possible d'intégrer toutes ces informations pour avoir une vision d'ensemble. Le développement floral peut être vu comme une évolution d'états moléculaires. Dans son exposé, Françoise insiste que le fait que le réseau a une structure universelle et un comportement dans chaque cellule, qui s'observe au travers du phénotype. Pour inférer le réseau, il y a eu (i) création d'une base de données avec les interactions moléculaires directes (données booléennes), des évidences d'induction et des interactions génétiques (travail sur les mutants) (ii) création d'une base de données d'expression à partir d'hybridation in situ et d'observation au microscope confocal. Ces deux bases ont été utilisées pour identifier 6 zones moléculaires à partir des données d'expression et proposer des réseaux candidats qui sont ensuite testés pour savoir si connaissances non utilisées pour les construire sont bien inférer dans ces réseaux candidats.

Frederik Gwinner nous a presenté sur son travail de thèse sur la modélisation de la réponse aux stress abiotiques (chaleur, sécheresse, salinité...) chez A. thaliana. Il utilise des données de cinétiques produites avec une puce Affymetrix. Frederik nous a présenté le modele Shiraishi et le « impulsion model ». Ce dernier modèle pose des questions d'estimation et de recalage des données. Frederik valide sur la base de données (nettoyée par du travail sur la littérature) AtRegNet (AGRIS) d'interactions directes TF vers cible de régulation.

Benno Schwikowski et Oriol Guitar ont ensuite présenté Cytoscape, qui permet de visualiser des réseaux. Une présentation et tutorial de Cytoscape sont disponibles (http://www.cytoscape.org/). Ils nous ont ensuite expliqué les grands changements qui vont avoir lieu dans la prochaine version (3.0) qui doit faciliter le travail des développeurs de l'équipe de Cytoscape mais également des développeurs qui proposent des plug-in. D'ailleurs les plugins deviennent des « Apps ».

Les journées se sont terminées par une discussion autour des réseaux et de la vie de NETBIO.

L'année dernière, la discussion avait principalement porté sur les verrous méthodologiques qui restait et sur la difficulté d'utilisation des méthodologies pour l'inférence de réseaux géniques. Le constat lors de ces nouvelles journées a été que nous avions maintenant une vision globale des méthodes existantes pour inférer un réseau uniquement à partir d'une seule source de données (transcriptomes en biologie) et que nous souhaitions nous tourner vers d'autres méthodes ou/et objectifs pour mieux inférer des réseaux car les performances des méthodes actuelles sont décevantes. Les thèmes qui ont émergés de manière consensuelle sont

- La causalité avec demande pour les prochaines journées d'inviter Peter Bühlmann (ETH Zurick).
- Prédire un phénotype en utilisant et/ou inférant les réseaux. L'ajout du phénotype semble compliquer le problème mais permettrait d'inférer un réseau qui a un objectif de fonctionnement pour l'organisme.
- S'investir dans la modélisation de données hétérogènes pour inférer un réseau a soulevé de nombreuses questions ou commentaires sur :

– la nécessité d'avoir un jeu de données propre et complet, y compris ds le réseau où il ne faut pas qu'il manque trop d'acteurs.

– l'objectif du travail. Souhaite-t-on travailler sur un réseau déjà connu pour réussir à trouver ce qui est déjà connu ou voulons nous un réseau « jeune » pour aussi proposer de nouvelles interactions. Les deux questions nous ont semblé pertinentes. Dans le premier cas, cela permettra de mieux comprendre les données nécessaires pour inférer un réseau et d'être capables ensuite de dire aux biologistes sous quelles conditions (données nécessaires, qualité, nature, type de résultats fournis, ...) il est possible d'inférer un réseau. On espère que ce genre de travail permettra de mieux comprendre pourquoi les

méthodes actuelles ont des performances aussi décevantes. Pour aborder la deuxième question, il faut déjà qu'on se rassure sur la première et répondre à cette question dans le cadre du collaboration identifiée pour comprendre de quoi auront besoin les biologistes pour valider de nouvelles interactions.

- Sur la nécessité d'une collaboration étroite avec des biologistes et bioinformaticiens pour ne pas faire n'importe quoi et pour avoir une connaisance précise des données.
- Sur la compatibilité des méthodes actuelles à l'analyse de données hétérogènes. Il faudra certainement proposer d'autres modèles.
- Sur le fait que le biologiste est généralement intéressé par l'activité de la protéine et que les ressources génomiques actuelles sont principalement sur les transcrits. La réponse des biologistes présents dans la salle est que c'est plus facile d'avoir accès aux transcriptomes et que cette affirmation n'est pas tout à fait vrai car le transcriptome est le premier niveau où l'on voit l'activité des gènes et qu'il est aussi intéressant à connaître. Ils nous ont dit aussi que le processus de construction du réseau est encore plus important que le réseau en soit.

Sur le fonctionnement, nous avons prévu de garder la réunion annuelle qui est source de discussions riches et de commencer des réunions plus thématiques pour les personnes intéressées. Les thèmes de l'année 2013 seront

- la visualisation qui peut aider à la compréhension des données mais aussi être un handicap pour proposer des modèles car on a tendance à interpréter au lieu de modéliser. Il n'empêche que la visualisation semble importante dans les réseaux pour pouvoir donner du sens biologique aux interactions.
- les forêts aléatoires pour mieux comprendre pourquoi leurs performances sont toujours meilleures que celles des autres méthodes.
- un travail autour du réseau présenté par Françoise Monéger pour l'utiliser comme un jeu benchmark pour essayer les méthodes existantes et comprendre leurs défauts. Travailler sur la modélisation pour lier avec le bibliome, la programmation d'ensembles-réponses ....

Cette année il y aura aussi une refonte du site web pour proposer un wiki collaboratif qui permettra à chacun de contribuer. Julien Chiquet s'est proposé pour faire le cadre.
L'idée est d'y mettre les différents travaux, les packages R ou autres programmes développés des participants de NETBIO, les rapports de stage et les thèses. L'objectif est de capitaliser la connaissance et de mieux se citer les uns les autres.

Une demande de financement du réseau sera soumise au département MIA de l'INRA, qui le finance depuis 2009. Mais nous avons aussi déposé une demande de financement au RNSC. Pour les prochaines demandes, nous solliciterons aussi les métaprogrammes INRA selgen et mem.
L'argent nous permettra de participer aux indemnités de stages, inviter des chercheurs sur de nouvelles approches, favoriser les échanges entre participants de NETBIO.