Compte-rendu des journées « Inférence de réseaux » qui ont eu lieu le 9 et 10 février à l'AgroParisTech.

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

Les journées ont rassemblé une trentaine de personnes le jeudi et une vingtaine le vendredi. 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

Marine Jeanmougin nous a présenté un travail permettant d'utiliser des informations biologiques extérieures *a priori* lors de l'inférence de réseaux à l'aide de modèles graphiques gaussiens. L'objectif est de réduire l'espace des réseaux candidats en pondérant la pénalité LASSO selon l'existence ou non d'information biologique sur la relation entre les gènes.

Claire Nédellec nous a présenté ses travaux sur les méthodes d'extraction de régulations géniques à partir d'articles. Les documents textuels sont une source majeure d'information en biologie encore peu exploitée. Depuis une dizaine d'années, les méthodes de fouille de textes dans ce domaine ont énormément progressé. Les co-citations, pourtant l'outil majoritairement utilisé pour la fouille de texte en bioinformatique, ne sont pas forcément un indice d'interaction génique : la moitié sont des faux-positifs. Claire a montré l'intérêt de faire une étape d'apprentissage pour mieux compter les relations géniques. Cela est fondé sur la création d'un dictionnaire et de traitements linguistiques *ad hoc*. Un outil informatique est aussi développé.

Sophie Lèbre nous a présenté une méthode d'inférence de réseaux à l'aide d'un modèle bayésien dynamique. L'originalité de la méthode tient dans une évolution possible des interactions au cours du temps. Cette méthode nécessite d'observer les gènes au cours d'une cinétique à pas de temps court. Dans la représentation graphique, chaque nœud est un gène à un temps donné. Ce travail s'appuie sur le papier de Audrieu et Doucet (1999) et cherche en plus à faire du partage d'information entre deux plages de temps successives. Ceci se fait à l'aide de prior bien choisis. L'estimation se fait à l'aide de MCMC.

Andrea Rau nous a présenté un travail sur les réseaux bayésiens pour retrouver la structure à l'aide d'une cinétique quand des variables cachées continues (gènes non observés ou mesures ou autres informations biologiques) existent. L'inférence du modèle par une méthode variationnelle est très rapide par rapport aux approches gourmandes en capacité de calcul comme les MCMC. Andrea nous a également montré que les résultats des analyses d'un même jeu de données par plusieurs méthodes étaient très peu cohérents.

Jimmy Vandel nous a présenté un travail sur un réseau bayésien statique discret où un *a priori* lié à la connectivité des réseaux est proposé pour réduire l'espace des réseaux candidats. Après une étude de simulation non présentée lors de l'exposé, Jimmy a montré les performances de la méthode pour inférer un réseau chez *Arabidopsis thaliana* à partir de données transcriptomiques sur des lignées recombinantes et des marqueurs génétiques. Les résultats ont été discutés au regard de données eQTL déjà disponibles.

Daniel Kahn nous a présenté le principe de l'analyse modulaire de réponse. A partir d'un système constitué de modules, on mesure la réponse de l'ensemble du système suite à une petite perturbation d'un module autour de son état stable (on perturbe tous les modules indépendamment). L'objectif est de caractériser le comportement des modules en fonction de celui des autres. La méthode est utilisable avec des siRNA et pas avec des expériences de *gene knock-out* (perturbation trop brutale).

Jean-Jacques Daudin nous a présenté une revue des méthodes de classification des nœuds d'un réseau pour identifier des modules indépendants dans un grand réseau. Deux types de module peuvent être trouvés (i) ceux dont les nœuds sont fortement connectés entre eux mais faiblement aux autres nœuds extérieurs au module (ii) ceux dont les nœuds ont des relations identiques dans et à l'extérieur du module (SES pour *structural equivalence of actors*). En fonction de la définition choisie, les modules trouvés seront différents mais apporteront des informations complémentaires. Jean-Jacques nous a ensuite présentés les différentes méthodes en les séparant en trois classes (i) les méthodes fondées sur un algorithme, (ii) les méthodes optimisant un critère et (iii) les méthodes probabilistes.

Antoine Channarond nous a présenté un travail sur le *Stochastic Block Model* (SBM) qui est une méthode probabiliste pour faire de la classification de nœuds dans un réseau donné. L'inférence dans ce modèle est délicate et est toujours faite à partir de méthodes itératives avec mises à jour successives de la classification des sommets et des estimateurs. Si la classification des nœuds dans les modules est faite, l'estimation est rapide. Antoine propose donc de travailler sur la distribution du degré normalisé des nœuds pour faire cette classification. Il a proposé un algorithme pour trouver des paquets de nœuds et prouvé que cet algorithme est consistant sous certaines conditions. Cet algorithme permet aussi de trouver le nombre de modules. Ce travail montre aussi que toute l'information des SBM est contenue asymptotiquement dans les degrés normalisés des nœuds.

Le vendredi après-midi a été consacré à une discussion autour de l'inférence de réseaux pour identifier de nouvelles questions méthodologiques qui intéressent les participants du réseau.
Les différents exposés avaient soulevés de nombreuses questions des participants et ont alimenté la discussion.
Il nous est apparu que la technicité liée à l'inférence d'un réseau est maintenant relativement bien maîtrisée même si de nombreux développements sont encore en cours ou envisagés. Par exemple liées aux problèmes de petite taille des d'échantillons, de temps de calcul parfois prohibitif, de détection de la causalité (*cf* les exposés de cette réunion et des réunions précédentes) : la prise en compte de telles particularité amène des défis aussi bien mathématiques qu'algorithmiques. Toutefois, l'application des méthodes d'inférence de réseaux soulèvent de nombreuses questions encore sans réponse.

- On constate généralement que la structure du réseau inféré est très dépendante de la méthode utilisée ce qui est un réel problème. Il est donc important de pouvoir donner un intervalle de crédibilité aux arêtes détectées.
- Nous constatons aussi que nous manquons de descripteurs (par ex., degré entrant/sortant d'un nœud, existence d'un chemin orienté ou non entre deux nœuds, modules, etc) pertinents d'un réseau car il semble que comparer la présence d'arêtes prédites en sortie de deux (ou plus) méthodes n'est pas suffisant.
- Il nous semble que cette « non-identifiabilité » du réseau ou ce manque de robustesse provient du fait que nous cherchons à inférer une structure très complexe dont tous les acteurs ne seront jamais observables (dans le cas de réseau de régulation biologique, les gènes sont importants mais également les phénomènes épigénétiques, les petits ARNs, les modifications post-transcriptionnelles/translationnelles...). Pour cela, nous avons identifié deux pistes de recherche :
  (i) inférer la structure à partir de données hétérogènes (Chip-seq, miRNAs, résultats de fouille de texte ...)
  (ii) savoir définir un sous-réseau pertinent pour inférer la structure recherchée dessus (cela résoudrait également le problème de temps de calcul).

Ainsi il nous apparaît que le cloisonnement des thématiques de recherche des membres de ce réseau aux seuls aspects techniques sur la développement d'une nouvelle méthode pour inférer un réseau à partir de données biologiques est désormais dépassé : le chercheur est amené à prendre en compte l'utilité des prédictions fournies et envisager des stratégies de validation variées de celles-ci.