Compte rendu de la réunion Netbio, 11/12 septembre 2013

AgroParisTech, Paris

Les journées se sont déroulées sur une journée et demi et ont rassemblé une quarantaine de personnes.

Mercredi 11 septembre : 13h-17h30

* Laurent Vallat présente un projet autour de la leucémie dont l'objectif est de mieux connaître le fonctionnement du programme génétique pour identifier les raisons pour lesquelles une cellule a une réponse inadaptée. Le but est de pouvoir, à terme, identifier des gènes sur lesquels agir pour proposer des solutions thérapeutiques plus ciblées. En conséquence, l'approche pour l'inférence n'est pas de modéliser la cascade intégrale d'événements provoqués par une stimulation exogène mais de chercher les régulations existantes et sur lesquelles on pourrait intervenir sans complètement modifiées l'ensemble du réseau de régulation. Les données collectées sont 136 observations de type données d'expression sur des cellules humaines (3 types de cellules : saines, indolentes, agressives pour 17 patients x 4 temps x 2 conditions). Les gènes sont classés de manière non supervisée selon leur profil d'expression temporel (fuzzy c-means) et la connaissance de l'information temporelle est intégrée dans la méthode d'inférence. Les résultats ont été évalués sur données simulées et comparées avec d'autres approches classiques (GeneNet, aracne, …) sur la base du F-score.
Nicolas Jung présente les aspects méthodologiques du travail. La méthodologie est basée sur une méthode qui repose sur des systèmes d'équations et utilise la régression LASSO. Un réseau en cascade est modélisé : les relations de régulation sont inférées en tenant compte du groupe temporel auquel le gène appartient et comporte trois étapes : sélection des gènes différentiellement exprimés ; construction des clusters temporels et régression linéaire avec pénalisation LASSO sur les différences d'expression et en utilisant les clusters temporels.

* Françoise Monéger présente un projet sur l'inférence de réseau lors du développement de la plante Arabidopsis. Elle présente une base de données contenant les interactions décrites dans 300 articles (directes ou indirectes) ainsi que des données d'expression correspondant à différents états de la cellule. De ces données est extrait un réseau bibliographique et ce réseau est validé en essayant de prédire, à partir du réseau, les différents états des cellules de la plante : les résultats montrent une bonne précision des prédictions à partir du réseau bibliographique.
Marie-Laure Martin-Magniette parle d'un prolongement de ce travail pour inférer le réseau à partir de données d'expression cette fois-ci. 4342 échantillons de données d'expression sont disponibles correspondant à 32 organes (dont 4 principaux), 82 stades de développement, 40 écotypes et 486 modalités de génotype.
Matthieu Vignes explique la méthodologie mise en œuvre pour l'inférence et qui est fondée sur 38 sondes annotées et qui font partie du réseau bibliographique présenté par Françoise (1558 expériences). La méthodologie utilise une approche par bootstrap : à chaque échantillon bootstrap étaient conservées les arêtes dont la corrélation partielle estimée était supérieure à la corrélation partielle d'un cas aléatoire. Les arêtes sélectionnées dans tous les échantillons bootstrap constituaient le réseau final : celui-ci contenait 70 arêtes qui incluaient les 50 arêtes du réseau initial.

Une discussion a ensuite été proposée pour échanger sur les deux exposés. Nous avons constaté que la notion de réseau est différente chez les biologistes et les modélisateurs. Le réseau biologique est le résumé de nombreux types d'interaction obtenus par de différentes approches. Ceci est très

différent des modèles graphiques proposés par les modélisateurs. Il a encore été souligné de la nécessité de déterminer si les modèles graphiques sont réellement adaptés pour répondre aux questions des biologistes. Il a été proposé de constituer un groupe de travail sur cette question en prenant le réseau de F. Monéger et les premières analyses réalisées comme fil conducteur.

Vous trouverez ci-dessous les différentes questions et constats que nous avons posées ou faits:

L'idée est de confronter les prédictions des modèles statistiques à la « réalité ».

- on pourrait ainsi dire si le transcriptome est la bonne données pour inférer telle ou telle relation.
- on pourrait aussi etudier les conséquences des hypothèses des modèle sur le réseau inféré. Par exemple
  - DAG: il y a souvent des circuits dans les réseaux biologiques et peu de méthodes permettent de les mettre en évidence.
  - Un GGM modélise corrélations partielles, sont-elles utiles ? Le modèle est-il pertinent par rapport à ce qui est à observer ds les données.
  - si un gène important est absent du modèle, quelle qualité du réseau inféré ?

Prendre un jeu de données et se poser les bonnes questions dessus.
Qu'arrive-t-on à inférer comme réseau à partir de données d'expression seulement ?
De quelles données a-t-on besoin pour retrouver un réseau en entier ?

Prédire versus valider des relations (sélection de variables ?)

Passer par des réseaux de co-expression ?

Bien trier les données en se concentrant sur certains phénotypes, sur certains tissus.

Faire de l'expérimentation pour valider des modèles et/ou des méthodes.

Jeudi 12 septembre : 9h00-18h00

* Julien Chiquet a fait une présentation des modèles graphiques gaussiens pour l'analyse de réseaux et a présenté plusieurs extensions de ce modèle pour intégrer des informations additionnelles. La première extension consiste à utiliser une information a priori sur la structure latente (non observée) du réseau en groupes denses via le « stochastic block model ». Une deuxième extension permet d'inférer des réseaux à partir de données d'expression observées dans différentes conditions en utilisant une approche jointe, soit via un mélange de matrices de covariance empiriques, soit via des pénalités de type group-LASSO. Enfin, une dernière extension s'attaque au problème des « réseaux multi-attributs » construit à partir de données de plusieurs natures (par exemple données d'expression et données protéomiques) : la matrice de covariance empirique est structurée en blocs et une pénalité de type LASSO permet d'estimer le modèle. Enfin, une méthode permettant de coupler l'analyse différentielle et l'inférence « fused-ANOVA » est évoquée et fera l'objet de la thèse de T. HA qui débutera en octobre 2013.

* Yuna Blum présente un travail qui s'inscrit dans le cadre d'une méthodologie de type « relevance network » (ou « co-expression network ») basée sur la correlation. Le travail s'intéresse à l'inférence d'un réseau en relation avec un phénotype extérieur. Le modèle est basé sur une analyse en facteur dans lequel des facteurs modélisent la variabilité commune à l'ensemble des gènes. Des informations de type « Gene Ontology » peuvent aussi être intégrées dans le modèle comme a priori biologique et de manière similaire, les corrélations partielles peuvent être estimées. Les méthodes

sont comparées sur un jeu de données de 338 gènes annotés et corrélés à la masse de graisse abdominale et montre que la méthode améliore la prédiction.

* Magali Champion a présenté des résultats théoriques sur un algorithme de type L2-boosting pour résoudre une régression linéaire multiple (plusieurs inputs) et multivariée (plusieurs output) : consistance et recouvrement du support. Ensuite, son exposé a présenté une reparamétrisation du problème de l'inférence d'un réseau via la matrice d'adjacence A=P.T.t(P) où P est une matrice de permutation (encode un ordre sur les éléments du réseau) et T une matrice strictement triangulaire inférieure (encode la topologie du réseau, pénalisée en L1 pour rester 'sparse'). Le problème n'est plus convexe, en particulier pour l'optimisation en P (passage par les matrices bi-stochastiques alors). Illustration sur des premières simulations simples.

* Nathalie Villa-Vialaneix présente une extension du GGM pour l'inférence de réseaux à partir de données d'expression obtenues dans des conditions expérimentales multiples. La méthode est basée sur une double pénalisation : une pénalité L1 assure la parcimonie des réseaux prédits tandis une pénalité de type L2 pousse l'estimation en direction d'un réseau consensuel. La méthode est combinée avec une approche par bootstrap et testée sur des données réelles et simulées.

* Néhémy Lim a présenté un travail permettant d'inférer des réseaux de régulation à partir de données temporelles. Les données temporelles sont utilisées via un modèle auto-régressif non linéaire qui est formalisé par une approche basée sur des noyaux à valeurs opérateurs. L'approche proposée, appelée OKVAR, peut être combinée à une approche par boosting (voir l'article « OKVAR-Boost » publié dans Bioinformatics en Juin 2013).

* Andrea Rau nous a parlé de l'utilisation de données d'intervention de type « knock-out » pour retrouver des relations causales entre gènes en les combinant à des données purement observationnelles. Ce travail se place dans le cadre de DAG. Des designs complexes de knock-out ont été réalisés (knock-out multiples ou bien partiels). La question d'un design 'optimal' i.e. apportant un maximum d'information en un minimum d'expériences. L'approche utilise un modèle linéaire combiné à une procédure d'ordonnancement des sommets et donne des résultats performants sur des données simulées (selon le modèle de réseau bayésien gaussien ou prise dans le challenge DREAM4).

* Pierre Hilson a présenté le projet Knowtator qui est un projet visant à annoter les relations phénotypes/génotypes dans le cadre du développement de la feuille d'Arabidopsis. Dans le cadre de ce projet, un grand nombre de papiers ont été lus et des annotateurs humains ont ajouté des données issues de 109 articles pré-sélectionnés en se restreignant aux données décrites dans le texte principal, dans la section Results et aux articles relatifs au développement de la feuille. L'information est structurée en Slot ("Plant part", "Process", "Localisation", "...) / Original text / Ontology et une surcouche a été développée comprenant une interface de requêtage et d'exploration et l'extraction de fichier XML avec référence à l'ID Pubmed.
On a ainsi une base de données structurée, traçable jusqu'aux sources, remplie par des experts qui pourrait permettre un traitement quantitatif automatique.

* Pierre Boissy nous a présenté un travail d'extraction automatique de textes pour créer un réseau de régulations. Ce travail se place dans le cadre de BioNLP, qui est une série de compétitions internationales (2009, 2011, 2013) dont le but est de comparer les méthodes sur des données standard "mesure". Les challenges de BioNLP se focalisent sur des relations complexes, finement attachées au texte. Pour chaque tâche les participants ont accès en ligne à la spécification de tous les événements avec leurs relations et les arguments valides, les données d'entraînement de développement et de test, les ressources linguistiques utiles, les programmes d'évaluation et d'analyse des erreurs. Le travail s'est focalisé sur un réseau de régulation de gènes chez la bactérie

pour lesquelles il y a peu de données d'interaction génique structurées.

Conclusions

Journées très riches, l'expérience va se poursuivre avec un quatrième animateur dans le réseau, Julien Chiquet. Si un nouvel appel MIA est proposé, un nouveau dossier sera posé. RNSC peut-être (lourd pour peu de soutien financier). Matthieu parle du projet VERINET (A. de la Fuente) dans le cadre du réseau COST, lié thématiquement à NETBIO, qui s'intéresse aux méthodes d'évaluation de l'inférence de réseaux en allant vers des données réelles.