Compte-rendu de la réunion NETBIO du mardi 25 novembre

Présents : Patrick Meyer, Nathalie Villa-Vialaneix, Mélina Gallopin, Trung Ha, Marie-Laure Martin-Magniette

Centres d'intérêt

- Amélioration d'un organisme (phénotype ou un métabolite) en modifiant les gènes. Les modèles éco-physiologiques permettent de lier le phénotype à des paramètres environnementaux et des paramètres métaboliques. Modèles possibles à développer: expliquer le réseau métabolique par une fonction du réseau de gènes.

- Retour pratique sur l'utilisation de méthodes pour faire de l'inférence de réseaux. Des retours sur le package huge.

Patrick : package R en phase de soumission avec des benchmarks pour comparaison de méthodes d'inférence

Patrick parle d'un package R à venir pour faire de la comparaison de méthodes d'inférence. Le package incorpore un ensemble de données simulées à partir d'une matrice d'adjacence, d'équations différentielles et de gestion du bruit pour produire des données d'expression. Cela permet de générer des réseaux de milliers de gènes décrits par une centaine d'expressions. Le package permet de comparer une douzaine de méthodes (pour l'instant). Type d'évaluation du réseau évalué : precision-recall sur les 20 % des meilleures arêtes parmi les n(n-1)/2 arêtes d'un graphe complet de n sommets

Nathalie : compte-rendu sur des articles sur l'inférence de réseaux

Partant d'un GGM, l'objectif est d'estimer les coefficients non nuls de la matrice d'adjacence.
Le nombre de variables p étant grand devant le nombre d'observations, n, on réécrit la question comme la recherche des coefficients non nuls dans p régressions dans un cadre de grande dimension.
Au cours de la discussion, on s'est demandé si la constante de la pénalité est la même sur toutes les p régressions ou si cette constante peut être varier entre les p régressions et quelles sont les conséquences et aussi l'interprétation des deux possibilités.

Nathalie a lu plusieurs articles sur le choix du paramétrage. La question est de savoir comment calibrer le nombre d'arêtes dans un graphe. Il y a de nombreuses méthodes pour sélectionner les paramètres des méthodes pour déterminer le nombre total d'arête pertinent (de manière classique dans le cadre linéaire, la CV ou BIC, des approches plus spécifique en sélection de variables comme BIC ou des approches dédiées à l'inférence de réseau comme extended BIC, PIC, basée sur une approche par permutation ou StARS basée sur du ré-échantillonnage, implémentées dans huge). La discussion tourne autour de la pertinence de la sélection du nombre d'arêtes versus la recherche d'une bonne précision sur les premières arêtes prédites : le problème est que la précision nécessite une vérité terrain ou a minima un gold standard. Si une telle vérité existe, alors le modèle peut être calibré à partir de celle-ci (mais on est dans un cadre de comparaison de méthodes alors) mais si elle n'existe pas, il existe parfois de l'information a priori, très partielle qui pourrait être utilisée. La question de la manière d'incorporer cette information dans les critères de sélection du réseau actuels est évoquée avant la fin de la réunion.