Apprendre un système stochastique complexe partiellement observé : application à la reconstruction de réseaux de gènes.

Responsables du stage

Victor Picheny (victor.picheny@toulouse.inra.fr, 05 61 28 54 39)
Matthieu Vignes (matthieu.vignes@toulouse.inra.fr, 05 61 28 57 41)

Durée de stage

4 à 6 mois (rémunérés env. 400 euros net/mois selon convention de stage en vigueur dans le laboratoire) . Idéalement, le stage débuterait en février 2014.

Mots-clés

Modèles probabilistes graphiques, données manquantes

Contexte

Unité d'accueil - L'unité MIA-T (Mathématiques et Informatique Appliquées à Toulouse) de l'INRA d'Auzeville est majoritairement constituée d'informaticiens et de mathématiciens ; elle étudie notamment les relations entre les entités qui composent un système complexe, la plupart du temps dans le domaine appliqué des organismes biologiques, même si beaucoup de ses travaux sont méthodologiques (méthodes statistiques, implémentations algorithmiques).

Motivation - Nous développerons l'exemple des réseaux de reconstruction de gènes.
Les gènes sont des informations héréditaires portées par les organismes. Après des étapes de transcription puis traduction, ils permettent la synthèse de protéines, véritables molécules actrices du monde vivant : elles permettent à l'organisme de se reproduire, d'assimiler et de transformer de l'énergie, et de se modifier pour s'adapter à son environnement. Bien que chacune des cellules contienne toute l'information génétique au sein de l'ADN, toute cette information n'est pas exprimée à tous les instants. Cette expression est régulée selon un programme (codé dans l'ADN) dépendant des conditions environnementales et de développement de l'organisme.
Aujourd'hui, des données dites "omiques, à haut débit", permettent des mesures instantanées qui reflètent l'activité des entités biologiques qui forment le modèle que nous avons de ces organismes : des systèmes complexes dont les composants sont ces entités et qui dialoguent entre elles. Souvent, ces données sont utilisées pour reconstruire les relations de régulation (c'est à dire de dépendance et d'indépendance) inconnues entre les composants du système. Les modèles graphiques (*e.g.* réseaux bayésiens) fournissent un cadre pertinent pour l'analyse statistique de ces données. Ils permettent de prendre en compte le caractère bruité, les soucis de normalisation, la haute dimension de ces données.

Cependant, ces données omiques sont des mesures imparfaites, partielles de l'activité des composants du système. Dans le cadre d'un réseau de régulation de gènes, seule une partie de ceux-

ci peut être effectivement mesurés, ou bien des acteurs (qui ne sont pas des gènes) ayant un impact fort sur le système ne sont pas observés, sans que leur nombre et leur effet ne soit réellement connu. Peu de travaux abordent cette question pourtant cruciale.

Questions de recherche - La problématique générique que nous souhaitons explorer est alors celle de la qualité de la représentation sous forme de graphe du système complexe incomplet par rapport à la modélisation graphique du système complet.
En particulier, on cherchera à répondre aux deux questions suivantes :
(i) Si on enlève des nœuds à un graphe, quelle est la structure de graphe entre les nœuds restants qui "ressemble" le plus au graphe complet ?
(ii) Si on enlève des nœuds à un graphe, comment évolue l' "association" entre 2 nœuds restants vue l'information qu'apportait l'ensemble des observations sur le graphe complet (les données sont alors prises en compte).

Travail de stage

Le travail attendu se décompose en :

- une étude théorique, essentiellement bibliographique, pour trouver des mesures ah hoc de similarités/distances entre graphes,
- la construction d'une typologie des situation rencontrées lorsque des nœuds d'un graphe sont manquantes. Cela pourra se faire par une étude exhaustive de blocs élémentaires, qui une fois assemblés peuvent rendre compte de la plupart des systèmes complexes et
- une étude empirique pour implémenter et tester ces explorations sur des données simulées et réelles.

Compétences requises

Solides connaissances mathématiques/statistiques (en particulier en apprentissage) pour comprendre les modélisations mises en jeu.
Bonne connaissance d'un logiciel de calcul scientifique, préférablement R et/ou Matlab.
Aucune connaissance préalable en biologie n'est nécessaire mais un goût pour ce domaine appliqué (ou un autre) est souhaitable.
Capacité et organisation pour le travail et les interactions en équipe.
Anglais scientifique.