Analyse de l'impact de différentes approches d'apprentissage automatique sur le pouvoir prédictif des modèles résultants

Encadrant principal : Tony Ribeiro
post-doc à l'École Centrale de Nantes
IRCCyN
Mail : tonyribeiro.research@gmail.com
http ://tony.research.free.fr

Co-encadrant : Morgan Magnin
maître de conférences à l'École Centrale de Nantes
IRCCyN
Mail : morgan.magnin@irccyn.ec-nantes.fr
http ://www.irccyn.ec-nantes.fr/ magnin/

Contexte

Grâce au développement récent de technologies de mesure à haut débit telles que les puces à ADN, les biologistes réussissent à obtenir une grande quantité de profils d'expression génétique. Il devient alors crucial de pouvoir connecter les données et de construire un modèle prédictif du réseau de régulation génétique. Le travail proposé ici se positionne dans ce contexte.
LFIT [4, 5, 2] est un framework dédié à l'identification de la dynamique des systèmes types machine à états. A partir d'observations discrètes d'un système, LFIT permet de déterminer les influences locales de ses composants. Les observations en question sont les transitions d'états du système. Depuis ces observations, LFIT construit et raffine, transitions après transitions, un ensemble de règles logiques qui peut reproduire le comportement du système, ce que l'on appelle la dynamique du système. Cette méthode permet, entre autres, l'apprentissage de réseaux booléens et l'identification d'automates cellulaires. En bio-informatique, cette technique peut également être appliquée à l'identification d'un réseau de régulation génétique à partir d'observations obtenues par expérimentations en laboratoire.
Le point fort de la méthode est l'identification de pattern récurrents dans l'évolution des valeurs de différent composants. Grâce à la capture de ces corrélations, LFIT peut apprendre des règles qui reproduiront le comportement du système observé. C'est ce qui nous intéresse dans ce projet : la capacité de cette méthode à simuler le comportement du système appris, permettant la prédiction de l'évolution du système dans une situation non observée. La précision de cette prédiction est grandement dépendante des données d'apprentissage et de la qualité de la discrétisation de ces données. Les algorithmes du framework n'ont aucune tolérance aux perturbations ou bruits qui peuvent être présents dans les transitions qu'ils analyse. La gestion des imperfections des données se fait alors avant l'apprentissage, en **pré-processing**, par une discrétisation adapté des données et leur division en différent ensemble indépendant pour cross-validation, par exemple. Et après l'apprentissage, en **post-processing**, par une simple de-discrétisation des données ou par Jusqu'à maintenant la méthode à montrer des résultats satisfaisant concernant la prédiction sur les données de série temporelle du DREAM4 [3, 1]. Cependant, l'extraction des interactions réelle entre les composants reste un

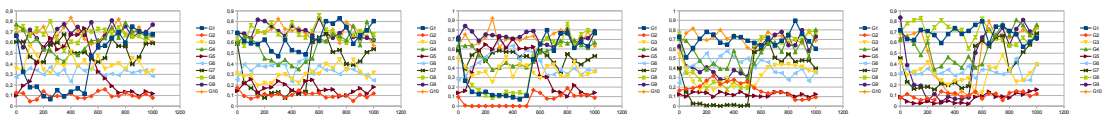

FIGURE 1 – Un des ensemble de série temporelle du DREAM4 pour un réseaux à 10 variables.

problème ouvert. En effet, les pattern qu'identifie LFIT peut concerner des composants qui n'ont pas d'interaction directe mais qui ont par exemple un rythme d'évolution concordant. C'est fort utile pour prédire l'évolution de ces composant lorsque que le système à un comportement habituel. Mais cela ne nous dit pas précisément quelles sont les interaction réelles, information utile lorsque l'on veut savoir comment réagira le système face à une perturbation forcée. Cependant, ces interactions réelles sont présentes dans les règles apprises par LFIT et devrais pouvoir être extraites d'une façon ou d'une autre. Et c'est là l'un des objectifs du stage.

Objectifs

L'objectif du travail proposé ici est d'améliorer les performance de prédiction du modèle appris par LFIT sur les données du DREAM4 challenge. Le champs d'action de l'étudiant se situe en amont et aval de l'algorithme d'apprentissage. Le gros du travail concernera la recherche et l'exploitation de particularité dans les données de série temporelle du problème qui constituent l'input de l'algorithme d'apprentissage. Cela peut inclure la détection de valeur aberrante dans les données, la caractérisation de perturbation, ce qui peut permettre d'affiner la méthode actuelle de discrétisation des données. Où même la proposition d'une méthode de discrétisation original dédiée à l'algorithme d'apprentissage utilisé dans ce projet.

Une deuxième partie concernera l'exploitation du modèle appris par LFIT combiné à d'autre données permettant de déterminer les relations réel entre les gènes. Le but ici sera de construire le graphe d'interaction des gènes : qui influence qui.

Environnement

L'équipe MeForBio (Méthodes Formelles pour la Bioiformatique) de l'IRCCyN (Institut de Recherche en Communications et Cybernétique de Nantes) est composée de trois permanents, un post-doctorant et cinq doctorants. Elle est impliquée dans plusieurs projets d'envergure nationale dont HyClock (ANR) et GRIOTE (projet fédérateur régional), et possède des partenariats nationaux et internationaux (notamment avec l'Allemagne et le Japon).

Travail à réaliser

- **compréhension du problème :** étude bibliographique initiale, permettant de comprendre les spécificités des données du DREAM challenge et des méthode d'apprentissage du framework d'apprentissage LFIT ;
- **exploitation des données :** implémentation de différentes méthodes existantes de discrétisation pour les séries temporelles adaptées aux données traitées ;
- **optimisation :** adaptation de ces méthodes pour optimiser la précision de prédiction du modèle appris ;
- **exploitation du résultat :** extraction de graphes de relation inter-gènes depuis le modèle appris par LFIT.

Références

[1] A. Greenfield, A. Madar, H. Ostrer, and R. Bonneau. Dream4 : Combining genetic and dynamic information to identify biological networks and dynamical models. *PloS one*, 5(10) :e13397–e13397, 2010.

[2] D. Martínez, T. Ribeiro, K. Inoue, G. Alenyà, and C. Torras. Learning probabilistic action models from interpretation transitions. In *The Technical Communications of the 31st International Conference on Logic Programming (ICLP 2015)*, 2015. To appear (short paper) (`http://tony.research.free.fr/paper/ICLP2015.pdf`).

[3] R. J. Prill, J. Saez-Rodriguez, L. G. Alexopoulos, P. K. Sorger, and G. Stolovitzky. Crowdsourcing network inference : the dream predictive signaling network challenge. *Science signaling*, 4(189) :mr7, 2011.

[4] T. Ribeiro and K. Inoue. Learning prime implicant conditions from interpretation transition. In *The 24th International Conference on Inductive Logic Programming*, 2014. To appear (long paper) (`http://tony.research.free.fr/paper/ILP2014long.pdf`).

[5] T. Ribeiro, M. Magnin, K. Inoue, and C. Sakama. Learning delayed influences of biological systems. *Frontiers in Bioengineering and Biotechnology*, 2 :81, 2015.