Inférence de réseaux biologiques : un défi pour la fouille de données structurées

Florence d'Alché-Buc*

*INRIA-Saclay, LRI et IBISC, Université d'Evry
florence.dalche@ibisc.fr

1 Résumé

La réponse cellulaire d'un organisme vivant à un signal donné, hormone, stress ou médicament, met en jeu des mécanismes complexes d'interaction et de régulation entre les gènes, les ARN messagers, les protéines et d'autres éléments tels que les micro-ARNs. On parle de réseau d'interaction pour décrire l'ensemble des interactions possibles entre protéines et de réseau de régulation génique pour représenter un ensemble de régulations entre gènes. Identifier ces interactions et ces régulations ouvre la porte à une meilleure compréhension du vivant et permet d'envisager de mieux soigner par le biais du ciblage thérapeutique. Puisque les techniques expérimentales de mesure à grande échelle, récemment développées, fournissent des données d'observation de ces réseaux, ce problème d'identification de réseau, généralement appelé inférence de réseau en biologie des systèmes, s'inscrit dans le cadre général de la fouille de données et plus particulièrement de l'apprentissage artificiel. Voilà maintenant quelques années que cette problématique a été posée à notre communauté et durant lesquelles les échanges entre biologistes et informaticiens ont non seulement permis aux biologistes d'étoffer leurs boîtes à outils mais aussi aux informaticiens de concevoir de nouvelles méthodes de fouille de données.

En partant des deux problématiques distinctes que sont l'inférence de réseau d'interaction et l'inférence de réseau de régulation, je montrerai que ces deux tâches d'apprentissage posent, chacune de manière différente, la problématique de la prédiction de sorties structurées. L'inférence de réseau d'interaction entre protéines, vue comme un problème transductif de prédiction de liens, peut être résolue comme un problème d'apprentissage d'un noyau de sortie à partir d'un noyau d'entrée. L'inférence de réseau de régulation, impliquant la modélisation d'un système dynamique, peut être abordée par l'approximation parcimonieuse et structurée de fonctions à valeurs vectorielles. Je présenterai un ensemble de nouveaux outils de régression à sortie dans un espace de Hilbert, fondés sur des noyaux à valeur opérateur, qui fournissent d'excellents résultats en inférence de réseaux biologiques. Des expériences in silico sur des données artificielles, chez la levure du boulanger ou chez l'homme illustreront mes propos. En fin d'exposé, je tracerai quelques perspectives concernant les " nouveaux " défis dans le domaine de la bioinformatique et dans celui de la prédiction de sorties structurées.