25
SEP.
2008

de 11h à 12h

SÉMINAIRE

Outils statistiques pour la sélection de variables et l'intégration de données "omiques"

Kim-Anh Lê Cao
Institut de Mathématiques de Toulouse

Les récentes avancées biotechnologiques permettent maintenant de mesurer une énorme quantité de données biologiques de différentes sources (données génomiques, protéomiques, métabolomiques, phénotypiques), souvent caractérisées par un petit nombre d'échantillons ou d'observations. L'objectif de ce travail est de développer ou d'adapter des méthodes statistiques adéquates permettant d'analyser ces jeux de données de grande dimension, en proposant aux biologistes des outils efficaces pour sélectionner les variables les plus pertinentes. Dans un premier temps, nous nous intéressons spécifiquement aux données de transcriptome et à la sélection de gènes discriminants dans un cadre de classification supervisée. Puis, dans un autre contexte, nous cherchons à sélectionner des variables de types différents lors de la réconciliation (ou l'intégration) de deux tableaux de données omiques. Dans la première partie de ce travail, nous proposons une approche de type wrapper en agrégeant des méthodes de classification (CART, SVM) pour sélectionner des gènes discriminants une ou plusieurs conditions biologiques. Dans la deuxième partie, nous développons une approche PLS avec pénalisation lasso dite de type sparse car conduisant à un ensemble ``creux" de paramètres, permettant de sélectionner des sous-ensembles de variables conjointement mesurées sur les mêmes échantillons biologiques. Un cadre de régression, ou d'analyse canonique est proposé pour répondre spécifiquement à la question biologique. Nous évaluons chacune des approches proposées en les comparant sur de nombreux jeux de données réels à des méthodes similaires proposées dans la littérature, . Les critères statistiques usuels que nous appliquons sont souvent limités par le petit nombre d'échantillons. Par conséquent, nous nous efforçons de toujours combiner nos évaluations statistiques avec une interprétation biologique détaillée des résultats. Les approches que nous proposons sont facilement applicables et donnent des résultats très satisfaisants qui répondent aux attentes des biologistes. Mots clés : sélection de variables, classification, sparse PLS, algorithme stochastique, biologie intégrative