Pharmacogénomique prédictive : identification des gènes différentiellement exprimés

René Natowicz - Mars 2008

ESIEE-Paris - Projet de fin de troisième année -
département d'informatique

Les microarrays donnent une mesure indirecte de l'expression des gènes des cellules composant un tissu organique auquel on s'intéresse. Les microarrays « pangénomiques » donnent cette mesure pour un ensemble de plus de 22 000 gènes.

Les données d'expression suscitent un très fort intérêt en médecine car on espère pouvoir prédire pour chaque patient, et avec une grande précision, l'efficacité des traitements envisagés. C'est le domaine de la pharmacogénomique prédictive. L'objectif est la construction de prédicteurs de l'efficacité des traitements. Les entrées de tels prédicteurs sont les niveaux d'expression d'un ensemble de gènes bien choisis et, éventuellement, les valeurs d'un ensemble de données cliniques et biologiques. La sortie est une prédiction de l'efficacité du traitement.

Pour ce problème, la sélection d'un sous-ensemble pertinent de gènes est un point central. Au travers d'essais cliniques on dispose, pour un ensemble de patients, des niveaux d'expression des gènes ainsi que de la réponse au traitement. On veut en déduire un ensemble de gènes dont les niveaux d'expression permettront une prédiction fiable de la réponse, pour les cas à venir.

En informatique et mathématiques il s'agit un apprentissage supervisé. Les données d'expression ont une caractéristique particulière: le nombre de cas (ici, le nombre de patients) est très faible comparé au nombre de données. Typiquement, moins de 100 cas pour plus de 22 000 données d'expression. Une autre caractéristique, non spécifique aux données d'expression, est que ces données sont très bruitées.

L'objectif de ce stage est la compréhension des différentes méthodes d'identification des gènes différentiellement exprimés entre les classes de patients, leur sélection, et la combinaison de leurs expressions. Certaines de ces méthodes sont déjà programmées et disponibles dans des bibliothèques de programmes publiques. Les autres seront programmées, selon les cas, en langage Java ou en langage R (environnement d'analyse de données statistiques). Leurs résultats seront comparés et analysés sur des données publiques.

Remarque : pour ce stage, il est préférable d'avoir suivi l'unité optionnelle « Introduction à la bioinformatique » (IN312).

Collaboration : Roberto Incitti, Institut Mondor de médecine moléculaire, Créteil ; Roman Rouzier, Hôpital Tenon, Service de gynécologie, Paris ; Antonio Braga & al. UFMG, Computational intelligence laboratory, Brésil (coopération CAPES-COFECUB).