



APPLIED STATISTICS WORKSHOP

Eric DEPIEREUX,
FUNDP, Namur,
Belgium

"Challenges in the Analysis of High Throughput Biology Data with Application to Microchips Data"

March 19, 2010

16:00

Room : **c 115 (STAT)**

Abstract

Les microarrays à DNA ont rendu possible la mesure simultanée de l'expression de milliers de gènes. Cette technologie permet d'aborder l'étude des systèmes biologiques sous un nouvel angle. La mesure globale de l'expression des gènes dans des systèmes biologiques complexes tels que des tissus, des cellules ou des populations devrait permettre de mieux caractériser leur fonctionnement et leur réaction à des stimuli ou des traitements. Cette technique, et d'autres dans la perspective de la "High throughput biology" produisent plusieurs dizaines de milliers de variables pour quelques réplicats et pose un défi nouveau aux statistiques traditionnelles, conçues pour fonctionner avec plus de réplicats que de variables. En particulier, l'estimation de la variance et les problèmes de tests sont mis en question. Les benchmarks des nouvelles méthodes posent également un problème, car soit les données sont simulées et la variance originelle n'est pas nécessairement bien modélisée, soit les données réelles ne permettent de déterminer la sensibilité et la spécificité des méthodes n'est pas connue précisément. Les ROC, confrontées à un nombre de vrais négatifs largement supérieur à celui des vrais et faux positifs, sont écartées au-delà de leur pouvoir discriminant.

Lorsque la modification de l'expression d'un gène est statistiquement significative, il ne s'agit que d'une information qui doit encore être validée par divers arguments. Le nombre de candidats reste généralement trop élevé pour permettre une confirmation expérimentale systématique, d'autant que les vrais positifs ne sont pas nécessairement « les plus significatifs ». La convergence de prédictions concernant la relation de gènes entre eux (co-expression, interaction des protéines produites, implication dans une même voie métabolique...) forme un faisceau d'arguments pouvant diminuer drastiquement la fréquence de faux positifs et/ou de faux négatifs.

Une information génomique considérable est en cours d'accumulation dans les bases de données bioinformatiques. Elle représente potentiellement une source d'information essentielle. Couplée à une méthodologie statistique, elle peut servir de filtre pour identifier les gènes potentiellement intéressants, pour éliminer les gènes considérés étrangers à la problématique étudiée, ou pour établir des groupes de gènes dont la modification d'expression est testée globalement, avec pour conséquence une augmentation de puissance du test.

You are welcome to the coffee break before the seminar (room : c 105)

Visit our page at: <http://www.uclouvain.be/72906.html>