SOCIETE FRANÇAISE DE STATISTIQUE

Journal de la Société Française de Statistique

Vol. 152 No. 1 (2011)

Méthodes Statistiques pour l'analyse de données post-génomiques

Title: Statistical methods for post-genomic data

Stéphane Robin¹

Depuis quelques dizaines d'années, grâce à une succession d'avancées technologiques, la biologie a effectuée des progrès considérables dans la compréhension des mécanismes fondamentaux du fonctionnement de la cellule. Ces progrès concernent tous les organismes des bactéries aux organismes supérieurs (végétaux, animaux, homme). Ces avancées se sont accompagnées de nouvelles questions posées à différents niveaux d'organisation : intracellulaire, de la cellule, du tissu, de l'individu, de la population voire de l'écosystème.

Ces progrès s'accompagnent d'une production massive de données dont le volume est passé, par expérience, de quelques kilo-octets à plusieurs giga-octets en une dizaine d'année. Les nouvelles technologies de séquençage (NGS) donnent aujourd'hui accès à l'information au niveau de chaque nucléotide (le génome humain en contient plusieurs milliards), les expériences de "métagénomiques" ciblent simultanément les génomes de milliers d'espèces présentes dans un milieu et les techniques d'hybridation in situ produisent des images complexes permettant de mesurer des abondances moléculaires au niveau intracellulaire.

Ces avancées se sont accompagnées de nouvelles questions méthodologiques qui ont fortement mobilisé les champs de l'informatique et des mathématiques appliquées, donnant naissance à la bioinformatique. La statistique constitue une composante essentielle de ce domaine à tel point que les revues de bioinformatique (et certaines revues de biologie) intègrent aujourd'hui explicitement l'évaluation statistique des méthodes ou des résultats présentés dans le processus de relecture.

Les données issues des technologies modernes de la biologie moléculaire présentent toute une série de caractéristiques intéressantes pour la statistique du fait de leur dimension, de leur hétérogénéité ou encore de leurs structures atypiques. L'objectif de ce numéro spécial est donner un aperçu de la recherche dans le domaine. Les articles qui le composent couvrent des sujets variés mais n'épuisent évidemment pas la diversité des questions statistiques du domaine.

Les contributions d'Etienne Roquain et de Bernd Klaus & Korbinian Strimmer traitent de problèmes de test multiples qui ont été reposés par les analyses de données à haut débit dès la fin des années 90. Ce sujet reste actif notamment sur la prise en compte de la dépendance entre les tests. Les techniques de régularisation issues de l'apprentissage sont également mobilisées pour l'analyse de ces données du fait de leur dimension et c'est l'objet de l'article de Laurent Jacob. Kim-Anh Lê Cao & Caroline Le Gall présentent des techniques de sélections de variables permet-

UMR 518 AgroParisTech / INRA Appl. Math. Comput. Sc. E-mail: Stephane.Robin@agroparistech.fr

tant elles aussi de circonscrire la dimension des données et s'intéressent de plus à l'intégration de données de nature différentes (expressions de gènes et mesures phénotypiques).

Enfin, depuis quelques années, l'analyse des réseaux biologiques décrivant les interactions ou les régulations entre les composants présents dans la cellule est devenue l'un des sujets phares des statistiques appliquées à la biologie. Dans leur article, Marine Jeanmougin, Mickael Guedj & Christophe Ambroise s'intéressent à l'inférence de réseaux, c'est à dire à la recherche des liens de régulation existant entre les différents gènes d'un organisme. La revue présentée par Jean-Jacques Daudin porte elle sur des modèles visant à décrire la structure sous-jacente à la topologie d'un réseau d'interaction.

Nous espérons que ce numéro spécial permettra aux lecteurs du journal de la SFdS de mesurer la richesse des recherches actuelles en statistiques appliquées à la biologie; nous espérons aussi que cette lecture donnera envie à quelques uns d'entre eux d'y contribuer. Je rappelle à cette occasion l'existence des journées " Statistical Methods for Post-Genomic Data " qui constituent depuis plusieurs années le rendez-vous de la communauté des statisticiens impliqués dans ces thématiques. La prochaine édition aura lieu à Lyon les 26-27 janvier 2012.

Pour conclure, je tiens à remercier Philippe Besse, rédacteur en chef du journal de la SFdS, qui est à l'initiative de ce numéro spécial. Je tiens également à remercier les auteurs pour leur contribution, ainsi que les relecteurs anonymes pour la qualité de leur travail.