Méthodes statistiques de pointe pour l'analyse intégrative de données massives

Directeur de thèse: Benoit Liquet (LMAP)

En raison d'un déluge de données sans précédent, il y a de l'espoir pour des découvertes monumentales dans les domaines biologiques tels que la génomique, la neuro-imagerie et des disciplines connexes. Néanmoins, il y a des défis méthodologiques statistiques qui doivent être surmontés en raison de contraintes de calcul, des effets non-linéaires, des mesures répétées et d'autres aspects concernant l'hétérogénéité dans les grands ensembles de données. Ce projet de thèse met en avant une méthodologie innovante et originale pour développer des outils qui permettront de surmonter le défi des données de grande dimension dans les domaines biologiques et révéler ainsi les secrets cachés dans les grands ensembles de données. De nouveaux outils sont nécessaires pour traduire les informations obtenues à partir des systèmes complexes. Les propriétés structurales des données provenant de la génomique et de la neuro-imagerie seront exploitées pour créer des techniques d'analyses statistiques originales. Cette proposition de thèse porte sur le développement de méthodes de calcul intensif en statistique et appropriées pour les "BIG DATA" afin de répondre à des questions biomédicales et biologiques clés. Les développements méthodologiques proposées dans ce projet de recherche, se penchera sur les questions les plus difficiles soulevées dans le domaine de la biologie computationnelle et proposer des alternatives à la méthodologie classique: (1) Il y a plus de paramètres à estimer que d'individus. (2) Le type/nature des données est hétérogène. (3) La relation entre les variables est souvent complexe (par exemple, non-linéaire). (4) La relation entre les variables change au fil du temps dans des études longitudinales. Pour faire face à toutes ces contraintes, ce projet de thèse va intégrer un large éventail de méthodes statistiques modernes intensives et informatiques. Celles-ci comprennent, des méthodes de réduction de dimension basés sur des indices (SIR, PLS), les approches de parcimonie (par exemple, pénalité lasso) et les méthodes de sélection de variables (par exemple, des approches Bayésiennes de selection de variables). Ces nouveaux développements basés sur des algorithmes pointus et efficaces seront implémentés et testés à l'aide de structure parallèle tel que le processeur graphique (GPU) disponible au Mésocentre de Calcul Intensif Aquitain (MCIA). Les méthodes développées seront déployées et mises à disposition sous forme de packages R.

La nouvelle méthodologie permettra aux scientifiques d'explorer et d'analyser les données massives dans des domaines émergents tels que "la biologie des systèmes" et "l'imagerie génétique". Exemples d'applications possibles comprennent: l'analyse intégrative pour étudier la réponse immunitaire au vaccin dans le contexte du virus de l'immunodéficience humaine (VIH); l'analyse des facteurs de variabilité (incluant les informations génétique) de la spécialisation hémisphérique du cerveau; étude de l'héritabilité de la connectivité anatomique dans le cerveau.

Référence:

Liquet, B., Lafaye de Micheaux, P., Hejblum, B., and Thiébaut, R. *Group and Sparse Group Partial Least Square Approaches Applied in Genomics Context.* Bioinformatics (2015).

Liquet B., Bottolo L., Campanella G., Richardson S. and Chadeau-Hyam M., R2GUESS: GPU-based R package for Bayesian variable selection regression of multivariate responses. Journal of Journal of Statistical Software (In press 2016).

Liquet B., and Saracco J. *BIG-SIR a Sliced Inverse Regression Approach for Massive Data. Statistics and Its Interface (In Revision 2016).*