Identifier des biomarqueurs en métagénomique quantitative : la promesse du Big Data

Magali Berland*[1], Emmanuelle Le Chatelier[†] , Edi Prifti[‡] , Nicolas Pons , Anne-Sophie Alvarez , Ndeye Gaye[§] , Jean-Michel Batto[2], Vincent Ducrot , Kévin Juilly , Sébastien Monot , Thierry Goubier , Nahid Emad[3], and Dusko Ehrlich

[1]INRA (Metagenopolis) – Institut national de la recherche agronomique (INRA) : US1367 – Domaine de Vilvert, 78 352 Jouy-en-Josas Cedex, France, France

[2]INRA US MetaGenoPolis 1367 – Institut national de la recherche agronomique (INRA) : US1367 – INRA DOMAINE DE VILVERT Unité MGP - Bâtiment 325 78352 Jouy-en-Josas Cedex, France

[3]Prism - Laboratoire d'informatique [Versailles] – CNRS : UMR8144, Université de Versailles Saint-Quentin-en-Yvelines (UVSQ) – 45 avenue des Etats-Unis Office 315 78035 Versailles, France

Résumé

Par des approches de séquençage à très haut-débit, la métagénomique quantitative consiste à quantifier les gènes ou les espèces composant un écosystème complexe (par exemple la flore intestinale humaine). Cependant, l'analyse des données métagénomiques se confronte à un défi de taille : la masse de données sans précédent générée par les nouvelles technologies de séquençage nécessite des outils de traitement et d'analyse de données adaptés à plusieurs millions de variables.
La suite *MetaOMineR* est un ensemble de packages R permettant d'explorer de nombreuses questions cruciales en métagénomique quantitative et a été largement déployé dans plusieurs projets fondateurs du domaine [1-2]. Elle a permis d'identifier des biomarqueurs prometteurs sur le plan de la santé humaine et permis des avancées significatives dans la quête de compréhension de ces écosystèmes complexes [3].

Le package central de la suite, *momr*, implémente une série de fonctions pour analyser des matrices de comptage des gènes présents ou absents au sein d'une cohorte d'individus en fonction d'un catalogue de gènes de référence. Des fonctions de normalisation et de sous-échantillonnage permettent de réduire la variabilité technique entre les séquençages. De plus, la réduction du nombre de dimensions est possible grâce à l'implémentation de fonctions de clustering, de projection sur les espèces métagénomiques (MGS) [4], et diverses procédures de filtrage et de mesures de qualité du signal. Enfin, un ensemble de fonctions statistiques permettent d'identifier les gènes, MGS ou fonctions associées à un phénotype ou à une donnée clinique et à visualiser les différentes relations et données.

La taille des catalogues de gènes sans cesse croissante (3.3 millions en 2010, 10 millions en 2014) limite l'utilisation de ce package avec des méthodes calculatoires classiques (calcul

*Intervenant
[†]Auteur correspondant: emmanuelle.lechatelier@jouy.inra.fr
[‡]Auteur correspondant: e.prifti@ican-institute.org
[§]Auteur correspondant: nagaye@jouy.inra.fr

lent, voire impossible). R n'étant pas adapté aux données Big Data, dans le cadre du projet européen MACH (*Massive Calculation of Heterogeneous System*) [5], il a été proposé de construire une solution permettant de rendre possibles ces analyses. Ce consortium est en train de développer un compilateur 'R' avec un nouvel IDE associé ayant pour cible des architectures data-parallèles comme le GPU ou le Xeon PHI. En parallèle, nous avons proposé l'implémentation d'une bibliothèque (version C/Cuda) ayant pour cible des architectures de type GPU dans le but d'optimiser le package *momr*.
Dans ce poster, nous présenterons le package ainsi que sa nouvelle implémentation accélérée pour une application sur des données réelles.