Santé, Nutrition et Big data : optimisation de packages R

Ndeye Aram GAYE [1], Jean-Michel BATTO [1], Nahid EMAD [2]

[1] INRA US MetaGenoPolis 1367
INRA DOMAINE DE VILVERT Unité MGP, Bâtiment 325, 78352, Jouy en Josas
nagaye@jouy.inra.fr, batto@jouy.inra.fr

[2] Laboratoire PRISM UMR 8144
Université de Versailles, Saint-Quentin en Yvelines, Versailles
emad@prism.uvsq.fr

Mots clés : métagénomique, Big Data, HPC, R, GPU, Xeon Phi

La santé humaine est fortement liée au microbiote humain. Il est difficile de l'étudier avec les méthodes classiques, c'est pourquoi la métagénomique s'est fortement développée ces dernières années.

Actuellement, beaucoup de données sont générées par la métagénomique: nous manipulons actuellement une matrice de 10 millions de gènes / 2000 individus, 20GB en terme de taille de fichiers et cela va passer à 40 millions de gènes pour un nombre d'individus donné. L'analyse de ces données est limitée par les méthodes computationnelles classiques (calcul lent voire impossible, outils inadaptés…). D'où la nécessité d'utiliser le HPC (en français Calcul Haute Performance) pour optimiser le processus d'analyse avec R.

C'est dans ce cadre que le projet ***MACH*** (projet ITEA2, ***www.mach-project.org***) a vu le jour. MACH pour Massive Calculation of Heteregous System est un projet européen regroupant notamment des équipes françaises, allemandes et belges. L'équipe française, constituée de l'***INRA*** (Institut National de recherche Agronomique), du ***CEA*** (Commissariat à l'Energie Atomique et aux Energies Alternatives) (www.cea.fr) et de la société de conseil et d'ingénierie ***AS+*** du groupe EOLEN (www.eolen.com), est chargée de développer un ***compilateur 'R'*** avec un nouvel IDE associé ayant pour cibles des architectures parallèles comme le ***GPU*** ou le ***Xeon PHI***.

Notre travail consiste à mettre en place un ***DSL*** (Domain Specific Language) appliqué à la métagénomique. Le but est d'optimiser nos temps de calcul et pour cela nous avons deux possibilités : développer soit un excellent compilateur (MACH) soit une librairie efficace.

Actuellement nous disposons d'un certain nombre de packages R, ***Megapack***, ***Parconnector***, ***GpuStat***, qui nous permettent d'effectuer nos calculs de manière simple et optimale et de passer d'une heure à deux heures de temps de calculs à quelques minutes.

Nous visons aussi les grilles de calculs avec des environnements qui intègrent un langage d'expression de parallélisme tel que YML (***yml.prism.uvsq.fr***).