OPTIMISATION D'UN PIPELINE POUR LA RECHERCHE DE BIOMARQUEURS DANS LE MICROBIOTE INTESTINAL

Spécialité : Informatique Haute Performance
Durée : 6 mois à partir de mars 2017, rémunéré par gratification.
Encadrement : Florence Thirion (florence.thirion@inra.fr, 01 74 07 16 31), Magali Berland (magali.berland@.inra.fr, 01 34 65 16 37)
Équipe d'accueil : MetaGenoPolis (US 1367), Bât. 325, INRA, Centre de recherche de Jouy-en-Josas, Domaine de Vilvert, 78350 Jouy-en-Josas.

CONTEXTE ET OBJECTIF DU STAGE

L'essor récent de la métagénomique a permis des avancées scientifiques majeures dans le domaine de la santé humaine en étudiant rôle du microbiote intestinal dans des maladies aussi diverses que l'obésité, le diabète, le cancer ou les maladies cardio-vasculaires.

Au sein de l'unité MetaGenoPolis, l'équipe InfoBioStat développe une expertise pointue dans la découverte et la caractérisation fonctionnelle de biomarqueurs microbiens associés à des pathologies. Dans ce contexte, IBS développe des solutions méthodologiques optimisées pour ces types de recherche.

Nous avons développé avec R et Shiny un pipeline de traitement statistique de données métagénomiques qui permet actuellement de traiter au maximum 200 échantillons (R ne gère plus la mémoire pour un nombre d'échantillons supérieur). Cependant, nous faisons face à une augmentation rapide du volume des données et les projets que nous avons à traiter vont maintenant au-delà (jusqu'à plusieurs milliers d'échantillons). D'autre part, la durée du temps de calcul (environ 10 heures) est également une limite aux analyses car il est courant de devoir exécuter le pipeline plusieurs fois (par exemple pour trouver les bons paramètres).

L'objectif ce de stage est de porter les points chauds du programme d'un langage interprété (R) à un langage compilé (C/C++) en utilisant au mieux les ressources disponibles (parallélisation, vectorisation). Dans un second temps, on étudiera la possibilité de distribuer le calcul sur plusieurs serveurs.

VOS MISSIONS

- Étudier, analyser et « profiler » le code existant (R),
- Proposer des méthodes d'accélération du code et de gestion de mémoire en prenant en compte les besoins des utilisateurs et l'exigence de maintenabilité du code pour améliorer la scalabilité
- Choisir la méthode à mettre en œuvre la plus adaptée et se former, si nécessaire, aux outils que requiert cette méthode (par exemple R avancé),
- Mettre en œuvre la méthode et la « benchmarker ».

Vous aurez à votre disposition :
Une infrastructure hautement performante avec un réseau à 10 Gbit/s et une cinquantaine de serveurs spécialisés pour les calculs, dont certains sont équipés de GPUs et d'accélérateur Intel Xeon Phi.

PROFIL SOUHAITE

- Formation M2 en informatique haute performance
- Connaissance du langage R
- Maitrise du C++
- Anglais courant