Défis computationnels des séquençage et phénotypage haut débit en science de la vie

Eric Rivals

Défis computationnels des séquençage et phénotypage haut débit en science de la vie

Eric Rivals - LIRMM, CNRS & Univ. Montpellier (rivals@lirmm.fr)

7 nov. 2014

La biologie et ses applications, de la médecine à l'agronomie ou l'écologie, deviennent des sciences productrices des données massives et par là exigent des approches computationnelles pour analyser ses données. Les nouvelles technologies de Séquençage à Haut Débit (SHD) apparues en 2005 révolutionnent la manière dont sont posées et résolues les questions de recherches en science du vivant. Par exemple, pour évaluer la biodiversité d'un espace, au lieu de déterminer patiemment les espèces après prélèvement, on peut aujourd'hui séquencer l'ADN des espèces présentes ou ayant laissé des traces dans un échantillon « environnemental » (sol, eau, air, intestin, etc). Une seule expérience de séquençage (ici de type métagénomique) produit plusieurs centaines de millions de courtes séquences, appelées lectures. Ces reads sont ensuite groupés en catégories représentant les espèces, et ainsi leur nombre et abondance relative permettent d'estimer la biodiversité. La question devient alors computationnelle. De même, l'accumulation de données structurées, de documents décrivant les procédures et résultats scientifiques de l'analyse des phénotypes du vivant requièrent des procédures informatiques pour exploiter et fouiller ces montagnes de données hétérogènes et réparties sur de multiple sites physiques connectés par des réseaux. Notre projet SePhHaDe cherche à proposer des solutions novatrices pour

- analyser des données massives de Séquençage à Haut Débit, les indexer et en extraire des informations biologiques,
- extraire des informations par requête d'un corpus de données distribuées, obtenues par divers plateformes sur les phénotypes de plantes, et effectuer de la recommandation automatique au bénéfice de l'utilisateur final.

Nos angles d'approche convoquent des techniques d'algorithmique du texte, d'indexation des données, de recherche d'information, ainsi que des techniques du oueb et des bases de données réparties. Je présenterai les enjeux des défis pour l'analyse des données du vivant, ainsi que des exemples de solutions proposées.

Pour plus d'information sur ce projet voir http://www.lirmm.fr/mastodons/ ; il est coordonné avec Esther Paccitti du LIRMM.