Proposition de stage de Master 2 Info 2018-2019

Encadrement: Jacques Nicolas, DR INRIA, Genscale. **jacques.nicolas@inria.fr**

Catégorisation de données grande échelle : application à la métagénomique

Mots-clés : Binning, Big data, Regroupement, Détection de communautés, Graphes, Bioinformatique.
Parcours visés : P4, P5
Lieu du stage : Equipe GENSCALE, IRISA, Rennes

Contexte : De nombreux domaines scientifiques ou industriels reposent maintenant sur la production d'importants volumes de données dont l'exploitation fait appel à de multiples méthodes regroupées sous le terme de « science des données » (data science). Une phase importante de l'analyse de telles données est leur regroupement préalable en entités cohérentes. Nous nous intéressons ici à l'étude de la biodiversité à partir de données métagénomiques, c'est-à-dire d'un mélange complexe de fragments d'ADN reflétant la composition de communautés bactériennes dans un environnement donné. On peut ainsi étudier aussi bien la flore intestinale humaine, la composition en micro-organismes d'un sol ou d'autres milieux comme l'air ou l'océan. Le but est de rassembler au mieux les fragments pour chaque espèce présente et ainsi à terme reconstituer le puzzle de leur génome.

Problème : D'un point de vue informatique/mathématique, on dispose d'un grand ensemble de séquences qui vont partager certaines caractéristiques communes pour une même espèce, ce qui va permettre de les regrouper. C'est un véritable challenge car un simple échantillon peut contenir des dizaines de millions de séquences. De nombreuses méthodes de regroupement (binning en anglais) ont été proposées depuis quelques années. Elles se basent principalement sur le développement d'indices de similarité entre jeux de séquences comme la couverture (le nombre de fois où on observe une même séquence) et la composition en k-mers (distribution des mots de taille k à l'intérieur de chaque séquence) [1], [2], auxquelles s'ajoutent parfois des connaissances plus spécifiques à la biologie comme le coalignement avec des espèces dont le génome est connu [3], l'usage des codons, caractéristiques des acides aminés utilisés par les différents organismes vivants [4] ou encore des marqueurs de la présence de gènes [5]. L'utilisation des critères peut être appliquée de façon hiérarchique [6], mais du côté algorithmique les méthodes sont généralement classiques (K-means, clustering hiérarchique, cartes auto-adaptatives de Kohonen…) et c'est le point d'amélioration que se propose d'étudier ce stage. Nous suggérons de s'inspirer pour cela de travaux réalisés pour l'identification de communautés d'intérêt qu'on retrouve par exemple pour analyser les données Web. L'idée majeure est de travailler globalement sur le graphe des similarités entre fragments et de former les communautés par une analyse topologique du graphe. Un premier travail récent propose ainsi de partitionner un tel graphe pour faire du binning [4]. Dans la thèse de C. Marchet au sein de l'équipe a été proposé un algorithme et un outil dans un cadre différent un peu plus simple [7], qu'il s'agirait d'adapter et d'étendre à ce nouveau contexte.

Expérimentation : Les algorithmes seront validés sur des jeux de données réels sur lesquels d'autres méthodes ont été testées [8] et pour lesquels on dispose également d'une expertise humaine montrant que l'approche manuelle est encore supérieure actuellement, au prix d'un travail très chronophage. On pourra utiliser à cette étape des outils de visualisation et d'évaluation des résultats[9, 10]. L'étude s'effectuera en coopération avec un spécialiste d'écologie microbienne environnementale, utilisateur des méthodes de binning, T. Delmont, qui travaille actuellement au département Bioinformatique du Centre National de Séquençage (Genoscope).

Références

[1] Alneberg, Johannes, et al. "Binning metagenomic contigs by coverage and composition." Nature methods vol 11 n°11 (2014): 1144.

[2] Herath, Damayanthi, et al. "CoMet: a workflow using contig coverage and composition for binning a metagenomic sample with high precision." *BMC bioinformatics* vol 18 n°16 (2017): 571.

[3] Lu, Yang Young, et al. "COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge." *Bioinformatics* vol 33 n°6 (2017): 791-798.

[4] Yu, Guoxian, et al. "BMC3C: Binning Metagenomic Contigs using Codon usage, sequence Composition and read Coverage." *Bioinformatics* vol 1 (2018): 8.

[5] Lin, Hsin-Hung, and Yu-Chieh Liao. "Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes." *Scientific reports* 6 (2016): 24175.

[6] Wang, Ying, et al. "Improving contig binning of metagenomic data using d2S oligonucleotide frequency dissimilarity." *BMC bioinformatics* vol 18 n°1 (2017): 425.

[7] C Marchet, L Lecompte, C Da Silva, C Cruaud, J-M Aury, J Nicolas, P Peterlongo "Clustering de Novo by Gene of Long Reads from Transcriptomics Data". *Nucleic Acids Researc*h 2018, to appear.

[8] Sharon, Itai, et al. "Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization." *Genome research* 23.1 (2013): 111-120.

[9] Broeksema, Bertjan, et al. "ICoVeR–an interactive visualization tool for verification and refinement of metagenomic bins." *BMC bioinformatics* 18.1 (2017): 233.

[10] Meyer, Fernando, et al. "AMBER: Assessment of Metagenome BinnERs." *GigaScience* (2018).