Titre du sujet : Outils de pan-génomique à l'épreuve des données réelles

Encadrant.e.s : Anne-Muriel Arigon (MAB, LIRMM), Sèverine Bérard et Sébastien Puechmaille (PEM, ISEM)

Équipes/Contacts : ISEM-PEM https://isem-evolution.fr/equipe/equipe-phylogenie-et-evolution-moleculaire/ et LIRMM-MAB https://www.lirmm.fr/equipes/MAB/
(Prénom.Nom@umontpellier.fr)

Mots clefs : Pan-génomes, structure de données, algorithmique, expérimentations

Résumé : (10 à 20 lignes avec des objectifs clairement définis)

L'information génétique, ou génome, de tout être vivant est codée sur des molécules d'ADN (ou d'ARN) et traitée de manière informatique sous la forme d'un texte écrit sur un alphabet à 4 lettres, généralement {A, C, G, T}. Les études biologiques/bioinformatiques au niveau des espèces se basaient jusqu'ici sur un unique génome de référence issu du matériel génétique d'individus de l'espèce en question. Les récents progrès dans les techniques de séquences permettent de nos jours d'avoir accès en temps et coût raisonnables au génome complet d'un individu. Nous avons alors plusieurs génomes, légèrement différents, pour une même espèce et la notion de référence unique devient caduque. Pour traiter toute la diversité intra-espèce, nous avons besoin de considérer en même temps tous les génomes d'une même espèce. Ainsi est née l'idée d'un pan-génome comme représentation de cette diversité. Plusieurs structure ont été proposées, qui ne sont plus des structures linéaires. Dans l'équipe PEM, nous disposons de jeux de données d'un champignon /Pseudogymnoascus destructans/ pour lesquels des pan-génomes pourraient être construits.

Les objectifs de ce stage sont :
* Comprendre la problématique de la pan-génomique [1,3]
* Étudier les différentes catégories de pan-génomes, en particulier elles basées sur les séquences [2]
* Recenser les outils existants de construction de pan-génomes [3]
* Sélectionner les outils pertinents et établir un protocole pour les tester sur les jeux de données de *Pseudogymnoascus destructans*
* Proposer et programmer des nouvelles fonctionnalités et/ou des améliorations au meilleur outil déterminé à l'étape précédente soit sous la forme d'un module à juxtaposer à cet outil ou bien en repartant d'une structure de données de pangénome existante et en implémentant directement les fonctionnalités souhaitées.

Bibliographie

[1] The Computational Pan-Genomics Consortium, Computational pan-genomics: status, promises and challenges, /Briefings in Bioinformatics/, Volume 19, Issue 1, January 2018, Pages 118–135, https://doi.org/10.1093/bib/bbw089

[2] Zekic, T., Holley, G., Stoye, J. (2018). Pan-Genome Storage and Analysis Techniques. In: Setubal, J., Stoye, J., Stadler, P. (eds) Comparative Genomics. Methods in Molecular Biology, vol 1704. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-7463-4_2

[3] Thèse de doctorat "Méthodes d'analyse comparative de la variabilité intraspécifique des pangénomes procaryotes" par Adelme Bazin, dirigée par Vallenet, David Microbiologie université Paris-Saclay 2022 http://www.theses.fr/2022UPASL008