Intitulé du stage : Machine learning et réseaux phylogénétiques pour l'inférence du pangénome du pommier

Laboratoires : **IRHS (Institut de Recherche en Horticulture et Semence)**
LAREMA (Laboratoire Angevin de Recherche en MAthématiques)

Encadrement : **Charles-Elie Rabier (Maitre de conférences Section Mathématiques et Applications)**
Kilian Raschel (Directeur de Recherches CNRS Section Mathématiques et Intéractions)

Contact : **Kilian.Raschel@math.cnrs.fr , charles-elie.rabier@univ-angers.fr**

Contexte :

Ce stage touche à la fois le pangénome et les réseaux phylogénétiques. Le pangénome est un sujet de grand intérêt à l'heure actuelle en biologie, et notamment le pangénome du pommier à l'IRHS. Les réseaux phylogénétiques constituent un domaine de recherche interdisciplinaire à l'interface entre mathématique, informatique et biologie. Des nombreux développements mathématiques sont possibles autour de ces réseaux (Ané et al., 2024 ; Xu et al., 2023 ; Allman et al., 2022 et 2023).
On proposera de nouvelles méthodes statistiques basées sur les graphes aléatoires, afin de décrypter le pangénome du pommier, modèle de choix à l'IRHS. Le stage se déroulera au sein de l'équipe « BioInformatics for plant DEFense Investigation » (BIDEFI) de l'IRHS, et de l'équipe « Analyse, Probabilité et Statistique » du LAREMA. Les thématiques investies dans ce stage sont les graphes aléatoires liés à la combinatoire, les processus stochastiques en évolution (coalescence, naissance et morts …), la statistique mathématique, la statistique computationnelle et l'analyse de données omiques.

Enjeux et état de l'art :

Les bouleversements impressionnants en biotechnologie ont généré un flot de données considérable permettant l'exploration systématique du vivant, à différentes échelles d'organisation et le recours croissant en biologie à la modélisation mathématique et aux sciences du numérique. La pangénomique (Sigaux 2000,Tettelin et al. 2005) vise une exploitation maximale des données : on ne se limite pas à un seul génome de référence, mais on considère une représentation de tout le contenu génomique d'une espèce (Durant et al, 2021).
Le pangénome comporte deux composantes : le « Core Genome » et le « Dispensable Genome ». Le « Core Genome » commun à tous les individus de l'espèce, se veut le génome minimal requis pour une cellule de vivre. D'après Tranchant-Dubreuil et al. (2019), le « Core Genome » est un ensemble commun de séquences partagé par tous les individus du groupe considéré, et se veut le génome minimal requis pour qu'une cellule vive. Le « Dispensable Genome » contient quant à lui, un grand nombre de séquences et un nombre surprenant de gènes (Monat et al., 2016). Chez les plantes, le « Core Genome » représente de 40 à 80 % de la totalité du pangénome. A titre d'exemple, le « Dispensable Genome » constitue respectivement 33.7%, 38.1% et 26% du pangénome chez le blé (Montenegro et al., 2017), chez le riz asiatique (Zhao et al., 2018) et chez la banane (Rijzaani et al., 2021).

Récemment, Wang et al. (2023) se sont intéressés au pangénome chez le pommier à partir de 13 accessions (4 sauvages, 9 cultivés) présentant une large diversité en terme de qualité du fruit et de résistance à la maladie. A noter que des séquences de poirier et de pêcher ont permis l'analyse comparative. Au final, 53803 familles de gènes ont été constituées et des différences significatives entre taille de familles de gènes chez le pommier ont été identifiées. A titre d'exemple, 183 familles de gènes ont connues des expansions notables, alors que 6 familles ont subi des réductions légères. Ces expansions et réductions significatives pourraient expliquer l'adaptation à de nouveaux environnements (cf. Wang et al., 2023). Les méthodes utilisées dans Wang et al. (2023) sont les suivantes. Les auteurs infèrent tout d'abord un arbre phylogénétique par les méthodes RaxML (Stamatakis, 2014) et IQTree, puis analyse l'évolution des familles de gènes par la méthode CAFE (De Bie et al, 2006). CAFE repose sur un arbre phylogénétique et sur le processus de naissance et de morts. L'arbre phylogénétique (i.e. arbre d'espèces) représente l'histoire évolutive globale des différentes espèces. Le processus de naissance et de morts évoluant à l'intérieur de l'arbre d'espèces représente pour chaque famille de gènes, les duplications et pertes de gènes. Ainsi, les familles de gènes sont de taille variables, et peuvent subir des expansions ou des réductions au sein des différentes espèces. Pour rappel, la taille d'une famille de gènes désigne le nombre de copies de gène dans chacune des espèces. L'arbre de gènes associé à cette famille est l'arbre aléatoire résultant du processus de naissance et de morts.

Il s'avère que les méthodes employées dans Wang et al. (2023) ne modélisent pas l'ensemble de phénomènes biologiques complexes et connus. Un enjeu actuel est de proposer des méthodes statistiques basées sur une modélisation de phénomènes biologiques désormais connus.

Objectifs du stage :

Dans ce stage, on modélisera simultanément a) l'histoire réticulée (e.g. hybridations) à travers les réseaux phylogénétiques, b) le tri de lignées incomplet via le « Multispecies network coalescent » (cf. Degnan, 2018), et c) les duplications et pertes grâce au processus de naissance et de morts. Les réseaux phylogénétiques (cf. Solis-Lemus et Ané, 2016) sont des graphes dirigés sans cycles dirigés et avec une seule racine, qui peuvent décrire les transferts horizontaux de gènes (e.g. bactéries), les phénomènes d'hybridations (e.g. plantes) ainsi que les introgressions (e.g.

plantes et animaux). De plus, le « Multispecies Network Coalescent » prend en compte à la fois le tri de lignées incomplet (ILS), l'évolution des séquences, et le fait qu'une lignée génétique puisse hériter du matériel génétique d'un de ces parents, avec une certaine probabilité (modèle de Yu et al., 2012).

Un premier objectif est de proposer une méthode statistique d'inférence de réseau phylogénétique reposant sur un modèle incluant duplications et pertes de gènes, mais aussi tri de lignées incomplet. On pourra s'intéresser à l'inférence Bayésienne en choisissant comme loi a priori sur le réseau phylogénétique un processus de naissance hybridation (Zhang et al., 2018). A noter que la statistique Bayésienne permet d'avoir accès à une distribution de réseaux, et d'ainsi de pouvoir quantifier l'incertitude sur certains clades (un clade est un groupe d'organismes comprenant un organisme particulier et la totalité de ses descendants). Une difficulté de ce travail réside dans l'estimation de la distribution a posteriori : la fonction de vraisemblance des données de séquences peut s'avérer complexe à calculer de manière analytique au vu des processus stochastiques investis (naissance et mort, ainsi que coalescence) évoluant à l'intérieur du réseau phylogénétique. Des méthodes Bayésiennes approchées (e.g. ABC-Random Forest cf. Pudlo et al., 2015) sont également envisageables. Dans le cadre de ce modèle, est-il possible d'intégrer sur tous les scénarios évolutifs (au sein du réseau) à l'instar de ce qui été proposé dans SnappNet (Rabier et al., 2021) uniquement dans le cadre du « Multispecies Network Coalescent » ?

Une fois le réseau phylogénétique inféré, un deuxième objectif est d'estimer, pour chaque famille de gènes, la meilleure reconciliation (i.e. plongement) de l'arbre de gènes dans le réseau phylogénétique. Cela permettrait pour chaque famille de gènes, de disposer de l'arbre de gènes reconcilié et d'ainsi de pouvoir mettre en évidence les gènes orthologues (issus d'une spéciation) des gènes paralogues (issus d'une duplication). Ceci s'avère primordial en génomique comparative, afin de mettre en relation les gènes qui assurent les mêmes fonctions, et s'attaquer au pangénome. Notre méthode d'inférence sera testée sur les données de familles de gènes de Wang et al. (2023). On sera ainsi en mesure d'affiner la figure 2 de Wang et al. (2023) : un réseau phylogénétique remplacera l'arbre d'espèce inféré, et les arbres de gènes remplaceront l'arbre résumant l'expansion et la réduction des tailles de familles de gènes.

Compétences recherchées :

-Statistique
-Processus stochastiques en évolution
-Bioinformatique

Références Bibliographiques :

Allman, E. S., Baños, H., & Rhodes, J. A. (2022). Identifiability of species network topologies from genomic sequences using the logDet distance. *Journal of mathematical biology*, *84*(5), 35.

Allman, E. S., Baños, H., Mitchell, J. D., & Rhodes, J. A. (2023). The tree of blobs of a species network: identifiability under the coalescent. *Journal of mathematical biology*, *86*(1), 10.

Ané, C., Fogg, J., Allman, E. S., Baños, H., & Rhodes, J. A. (2024). Anomalous networks under the multispecies coalescent: theory and prevalence. *Journal of Mathematical Biology*, *88*(3), 29.

Degnan, J. H. (2018). Modeling hybridization under the network multispecies coalescent. *Systematic biology,* 67, (5), 786-799.

De Bie, T. et al. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271.

Du, P., Ogilvie, H. A., & Nakhleh, L. (2019, May). Unifying gene duplication, loss, and coalescence on phylogenetic networks. In *International Symposium on Bioinformatics Research and Applications* (pp. 40-51). Cham: Springer International Publishing.

Durant, É., Sabot, F., Conte, M., Rouard, M. (2021). Panache: a Web Browser-Based Viewer for Linearized Pangenomes. *Bioinformatics*, 1-3.

Ebersberger, I., Galgoczy, P., Taudien, S., Taenzer, S., Platzer, M., & Von Haeseler, A. (2007). Mapping human genetic ancestry. *Molecular Biology and Evolution*, 24, (10), 2266-2276.

Mirarab, S., Nakhleh, L., & Warnow, T. (2021). Multispecies coalescent: theory and applications in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *52*, 247-268.

Monat, C., Pera, B., Ndjiondjop, M. N., Sow, M., Tranchant-Dubreuil, C., Bastianelli, L., ... & Sabot, F. (2016). De novo assemblies of three Oryza glaberrima accessions provide first insights about pan-genome of African rices. *Genome biology and evolution,* 9, (1), 1-6.

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C. K. K., ... & Edwards, D. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal*, 90, (5), 1007-1013.

Pudlo, P., Marin, J. M., Estoup, A., Cornuet, J. M., Gautier, M., & Robert, C. P. (2015).
Reliable ABC model choice via random forests. Bioinformatics, 32, (6), 859-866.

Rabier, C. E., Berry, V., Stoltz, M., Santos, J. D., Wang, W., Glaszmann, J. C., ... & Scornavacca, C. (2021). On the inference of complex phylogenetic networks by Markov Chain Monte-Carlo. *PLoS Computational Biology*, *17*(9), e1008380.

Rijzaani, H., Bayer, P. E., Rouard, M., Doležel, J., Batley, J., Edwards, D. (2021). The pangenome of banana highlights differences between genera and genomes. *The Plant Genome*.

Rokas, A., Williams, B.L., King, N., & Carroll, S.B. (2003). Genome scale approaches to resolve incongruence in molecular phylogenies. *Nature,* 425(6960), 798-804.

Sigaux, F. (2000). Cancer genome or the development of molecular portraits of tumors. *Bulletin de l'Academie nationale de medecine,*184, (7), 1441-7.

Solis-Lemus, C., Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *Plos Genetics*, 12(3), e1005896.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., ... & DeBoy, R. T. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial pan-genome. *Proceedings of the National Academy of Sciences*, 102(39), 13950-13955.

Tranchant-Dubreuil, C., Rouard, M., Sabot, F. (2019). Plant Pangenome : Impacts on Phenotypes and Evolution. *Annual Plant Reviews online*, 1-25.

Wang, T., Duan, S., Xu, C., Wang, Y., Zhang, X., Xu, X., ... & Wu, T. (2023). Pan-genome analysis of 13 Malus accessions reveals structural and sequence variations associated with fruit traits. *Nature Communications*, 14(1), 7377.
.
Xu, J., & Ané, C. (2023). Identifiability of local and global features of phylogenetic networks from average distances. *Journal of Mathematical Biology*, *86*(1), 12.

Yu, Y., Degnan, J. H., Nakhleh, L. (2012). The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genetics,* 8(4), e1002660.

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., ... & Wang, Y. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature genetics*. 50(2), 278.

Zhang, C., Ogilvie, H. A., Drummond, A. J., & Stadler, T. (2018). Bayesian inference of species networks from multilocus sequence data. *Molecular biology and evolution*, *35*(2), 504-517.