Topic : Machine learning and phylogenetic networks for inferring the apple pangenome

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

Mentoring : **Charles-Elie Rabier (Associate professor in Applied Mathematics)**
Kilian Raschel (CNRS Research director in Mathematics at CNRS)

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

Context :

This internship covers both the pangenome and the phylogenetic networks. The pangenome is a hot topic in biology, and specially the apple pangenome at IRHS. Phylogenetic networks are an interdisciplinary area of research, involving mathematics, computer science and biology. Many mathematical developments on phylogenetic networks are possible (Ané et al., 2024 ; Xu et al., 2023 ; Allman et al., 2022 et 2023). We will elaborate new statistical methods based on random graphs, in order to decipher the apple pangenome. This work will be within the team « BioInformatics for plant DEFense Investigation » (BIDEFI) from IRHS, and within the team « Analysis, Probability and Statistics » from LAREMA. We will tackle the following topics : random graphs linked to combinatorics, stochastic processes in evolutionary biology (coalescent process, birth and death …), mathematical statistics, computational statistics and omics data analysis.

Background :

Pangenomics (Sigaux 2000, Tettelin et al. 2005) aims to make maximum use of data: we do not focus on a single reference genome, but we consider a representation of the entire genomic content of a species (Durant et al, 2021). The pangenome has two components: the "Core Genome" and the "Dispensable Genome". The "Core Genome" common to all individuals of the species, is the minimum genome required for a cell to live. According to Tranchant-Dubreuil et al. (2019), the "Core Genome" is a common set of sequences shared by all individuals, and is intended to be the minimum genome required for a cell to live. The "Dispensable Genome" contains a large number of sequences and a surprising number of genes (Monat et al., 2016). In plants, the "Core Genome" represents 40 to 80% of the entire pangenome. For example, the "Dispensable Genome" constitutes 33.7%, 38.1% and 26% of the pangenome in wheat (Montenegro et al., 2017), Asian rice (Zhao et al., 2018) and bananas (Rijzaani et al., 2021), respectively.

Recently, Wang et al. (2023) investigated the apple pangenome thanks to 13 accessions (4 wild, 9 cultivated) with a wide diversity in terms of fruit quality and disease resistance. Note that sequences from pear and peach trees served as outgroups for the comparative analysis. Overall, 53803 gene families were build and significant differences between the size of gene families in apple trees were identified. For instance, 183 gene families experienced notable expansions, while 6 families underwent slight reductions. These significant expansions and reductions could explain adaptation to new environments (cf. Wang et al., 2023). The methods used in Wang et al. (2023) are as follows. The authors first infer a phylogenetic tree using RaxML (Stamatakis, 2014) and IQTree, and then analyze the evolution of gene families using CAFE (De Bie et al, 2006). CAFE is based on a phylogenetic tree and the birth and death process. The phylogenetic tree (i.e. species tree) represents the global evolutionary history of the different species. The birth and death process evolving within the species tree represents for each gene family, the gene duplications and gene losses. Thus, gene families are of variable size, and can undergo expansions or reductions within different species. Recall that the size of a gene family refers to the number of gene copies in each species. The gene tree associated with this gene family is the random tree resulting from the birth and death process.

It turns out that the methods used in Wang et al. (2023) do not model all the known biological phenomena. A current challenge is to propose sophisticated statistical methods based on a model taking into account recent progress in biology.

Research work :

During this internship, we will model simultaneously a) reticulated history (e.g. hybridizations) through the phylogenetic networks, b) incomplete lineage sorting through the multispecies network coalescent (cf. Degnan, 2018), and c) duplication and losses through the birth and death process. Phylogenetic networks (cf. Solis-Lemus et Ané, 2016) are directed acyclic graphs with a unique root : they can model horizontal gene transfer (e.g. bacteria), hybridizations (e.g. plants), and introgressions (e.g. plants and animals). Moreover, the Multispecies Network Coalescent takes into account the incomplete lineage sorting (ILS), the evolution of sequences, and the fact that a genetic lineage can inherit genetic material from one of these parents, with some probability (model by Yu et al., 2012).

A first objective is to propose a statistical method of phylogenetic network inference based on a model including gene duplications and losses, but also incomplete lineage sorting. Bayesian inference can be considered by choosing an hybridization birth process as a prior distribution on the phylogenetic network (Zhang et al., 2018). Note that Bayesian statistics give access to a distribution of networks : it enables to quantify the uncertainty on certain clades (a clade is a group of organisms comprising a particular organism and all of its descendants). A difficulty of this work lies in the estimation of the posterior distribution: the likelihood function of sequence data can be complex to compute analytically in view of the stochastic processes involved (birth and death, as well as coalescence) evolving within the phylogenetic network. Approximate Bayesian methods (e.g. ABC-Random Forest cf. Pudlo et al., 2015) could also be investigated. Within the framework of this model, is it possible to integrate over all evolutionary scenarios (within the network) as proposed in SnappNet (Rabier et al., 2021) only in the context of the "Multispecies Network Coalescent"?

Once the phylogenetic network has been inferred, a second objective is to estimate, for each gene family, the best reconciliation of the gene tree in the phylogenetic network. This would allow to differentiate, for each gene family, the orthologous genes (resulting from speciation) and the paralogous genes (resulting from duplication). This is key in comparative genomics, in order to link genes with the same functions, and to tackle the pangenome. Our inference method will be tested on gene family data from Wang et al. (2023). On Figure 2 of Wang et al. (2023), a phylogenetic network will replace the inferred species tree, and gene trees will replace the species tree that sums up expansion and reduction of gene family sizes.

Skills :

-Statistics
-Stochastic processes in evolutionary biology
-Bioinformatics

References :

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.