PHYLOGENOMICS AND POPULATION GENOMICS: MODDELS, ALGORITHMS, AND ANALYTICAL TOOLS

LUAY K. NAKHLEH
Department of Computer Science, Rice University, Houston, Texas, USA

NOAH A. ROSENBERG
Department of Biology, Stanford University, Stanford, California, USA

TANDY WARNOW
Department of Computer Science, University of Texas, Austin, Texas, USA

Advances in phylogenetics and population genetics have produced increasing awareness of the existence of problems of interest to both fields, and of problems for which approaches from one of the two areas can inform developments in the other. Phylogenetics and population genetics examine similar topics, but at different biological scales. Both fields analyze genetic similarities and differences among organisms, and the evolutionary processes that generate those similarities and differences. Whereas population genetics considers individuals and populations within species, phylogenetics focuses on relationships among species themselves. The two fields share a number of overlapping tools, as well as similar data in the form of biological sequences. Further, they come into direct contact in the analysis of populations that are sufficiently distantly related that they differ as much as distinct species, or species that are sufficiently closely related that they approach the level of population differences.

The increasing availability of genetic sequences within and among species has made the connections between phylogenetics and population genetics all the more apparent. *Phylogenomics* and *population genomics* have emerged as subjects concerned with phylogenetic and population-genetic problems at a genomic scale. The interface of phylogenetics, population genetics, and genomics has now generated significant research challenges, demanding new evolutionary *models* for linking population-genetic and phylogenetic time scales, new *algorithms* for analyzing data in contexts that involve both population-genetic and phylogenetic perspectives, and new *analytical tools* for assessing properties of the algorithms. We are pleased to present five papers that represent a range of topics that link phylogenomics and population genomics, and that demonstrate a range of ways in which models, algorithms, and analytical tools can be used to advance the subject.

The papers of James H. Degnan and Sebastien Roch address a traditional problem at the intersection of phylogenetics and population genetics: the modeling and inference of species phylogenies when incomplete lineage sorting generates discordance among gene trees. Roch takes a modeling approach to the comparison of several algorithms for species tree inference. In a three-species model, for each of the algorithms, Roch uses large-deviations theory to analytically determine the rate at which the probability of failing to infer the correct species tree decays for large numbers of loci. A higher decay rate indicates that a method has more favorable performance. The paper, which uncovers a substantial difference among methods,

produces an innovative analytical framework that can potentially be used for studying methodological performance in greater generality.

Whereas Roch presents a general perspective for assessing multiple methods, the complementary paper of Degnan instead looks closely at a single species tree inference method, the STAR approach (for Species Tree estimation using Average Ranks). In this method, the internal nodes of each of a series of input gene trees are given discrete ranks. For two species, the rank of their most recent common ancestor is averaged across all gene trees, and a species tree is inferred from the matrix of average ranks for all pairs of species. Degnan determines that the node-numbering scheme of the original STAR method is only one among many in a family of sensible schemes. By allowing variations of STAR that select from among this family of numbering schemes, Degnan finds that STAR can be generalized and enhanced. Together, the Degnan and Roch papers provide advances in the development and evaluation of new methods for inferring species trees in the presence of incomplete lineage sorting.

Two additional papers, one by Md. Shamsuzzoha Bayzid, Siavash Mirarab, and Tandy Warnow, and the other by Yu Lin, Fei Hu, Jijun Tang, and Bernard M. E. Moret, study species tree inference under another form of gene tree discordance. Gene trees can disagree not only because of incomplete lineage sorting, but also as a consequence of gene duplications and losses that lead to errors in assigning orthology across species. Duplications and losses begin by a population-genetic process: a macromutation arises, carrying a duplication or loss, and that mutation eventually spreads in a population. Once fixed, the duplication or loss becomes a phylogenetic character useful for analyses of species relationships.

Bayzid *et al.* algorithmically investigate the inference of species trees in the presence of discordance due to gene duplication and loss. They study a pair of optimization problems, one considering only duplications, and the other also considering losses. These problems are known to be NP-hard. The authors formulate solutions via a max (or min) clique problem, and they provide theoretical results for solving certain constrained versions of the problems. They obtain exact solutions for the constrained problems, providing algorithms that run polynomially in the number of genes and the number of taxa. The max clique formulation has similarities to work of Than & Nakhleh in the context of incomplete lineage sorting, thereby highlighting connections among algorithms for different processes that generate gene tree discordance.

Lin *et al.* consider genome rearrangements and insertions in addition to duplications and losses. They introduce a method for inferring species trees from genome sequences, accounting for the various types of gene content and gene order differences observed across species. Their method relies on an encoding of the spatial orientation of genes, and it performs inference with maximum likelihood. Lin *et al.* test their method using simulations incorporating a variety of factors, such as different simulated species trees, different combinations of evolutionary events along the tree, and errors in genome assembly. After obtaining favorable performance in their simulations, Lin *et al.* apply their approach to obtain a phylogeny of 68 genomes. As was true for the Degnan and Roch papers, the Bayzid *et al.* and Lin *et al.* papers illustrate two complementary components of the development and evaluation of phylogenomic algorithms, with Bayzid *et al.* proving theorems pertaining to the performance of their method, and Lin *et al.* examining simulations and an empirical assessment.

The final paper, by Naama M. Kopelman, Lewi Stone, Olivier Gascuel, and Noah A. Rosenberg, presents an example of a different aspect of the intersection of phylogenomics and population genomics. Kopelman *et al.* investigate the consequences of using a method borrowed from phylogenetics—the neighbor-joining algorithm—in a specific population-genetic context, namely that of admixed populations. Motivated by peculiar observations seen for admixed populations in neighbor-joining trees, they study neighbor-joining applied to populations that follow an admixture model. Kopelman *et al.* provide a mathematical demonstration under special cases of the model that admixed populations are expected to appear toward the middle of neighbor-joining trees, often with short branch lengths. The paper illustrates how mathematical analysis of a phylogenetic algorithm can be performed in a population-genetic setting.

Together, this collection of five papers highlights both the variety of algorithmic, mathematical, and statistical approaches now under development for investigating the interface of phylogenomics and population genomics, and the variety of problems of interest to both subjects. The proliferation of phylogenomic and population-genomic data will only increase the attention given to these problems, and the importance of devising sound models, algorithms, and analytical tools for addressing them is only likely to increase.

Acknowledgments

We gratefully acknowledge the support of National Science Foundation Collaborative Research grants DBI-1062335 (TW), DBI-1062463 (LKN), and DBI-1146722 (NAR).