Project title: Rethinking computational phylogenetics in the big data era

Murray Patterson (Assistant Professor, Dept. of Computer Science, Georgia State University, PI)

Project Summary. Since Darwin in the 1850s, the *de facto* model for representing evolutionary relationships between a set of objects is the *phylogenetic tree*, or *phylogeny*. More than a century later, computational methods for constructing phylogenies such as maximum parsimony (Fitch, 1971) and neighbour joining (Nei, 1987) appeared. By the 1980s, the advent of *molecular phylogenetics* had discarded Darwin's purely selectionist theory in favor of a more data-driven neutral theory of evolution (Kimura, 1983). As the field of molecular evolution continued to explode, new computational models and methods based on maximum likelihood (Felsenstein, 1978) and Bayesian approaches (Drummond, 2007) became the standard methods for constructing phylogenies.

Nowadays, with the decrease in costs of high-throughput sequencing, the amount of available data, *e.g.*, number of sequences, is orders of magnitude what it was a decade ago, and is accelerating quickly. Moreover, with advances in different sequencing technologies, the type of data being collected is becoming more diverse. The increasing *amount* puts a strain on the *scalability* of existing computational methods for reconstructing phylogenies, while the increasing *diversity* of types of data challenges the generality of current models for its representation. An example of the increasing amount is that the COVID-19 pandemic has made more than 15 million SARS-CoV-2 sequences available on databases such as GISAID (gisaid.org). State-of-the-art maximum likelihood methods such as Nextstrain (Hadfield, 2018) can build a tree on several thousand sequences, while IQTREE-2 (Minh, 2020) can scale to tens of thousands with parallelization. An example of the increasing diversity is that single-cell DNA sequencing allows us to sequence a cancerous tumor at the resolution of a single cell. At this resolution, there is widespread evidence of back-mutations (Kuipers, 2017), which in turn discards the neutral theory of evolution of Kimura (1983). Current approaches for reconstructing tumor phylogenies fall back to the simple parsimony methods, since so little is known about this environment that the assumptions needed for using maximum likelihood and Bayesian approaches would be arbitrary, and, most likely, more damaging than making no assumption to begin with (Kolaczkowski, 2004).

The goal of this proposal is to rethink computational phylogenetics from both a *modeling* and *methods* perspective, in working towards developing a general framework which can both scale to the current *amount* of data, but also to accurately model the increasingly *diverse* types of (sequencing) data. This will involve new algorithms and artificial intelligence approaches for coping with this "big data" challenge.

Intellectual Merit. The proposal aims to understand the modeling and methods challenges of larger and more diverse data from a theoretical point of view, and to devise a general approach to the problem. While there exists a handful of approaches to building, *e.g.*, SARS-CoV-2 phylogenies (O'Toole, 2021), which can scale to millions of sequences, they tend to be ad-hoc for SARS-CoV-2, and not for building large phylogenies in general. While there exist dozens of tools for inferring tumor phylogenies from single-cell DNA sequencing data, many of them are also ah-hoc, and only work for one tumor type, or (even worse) one ad-hoc simulated dataset. This proposal will build upon some of the techniques developed by the PI and his team, such as classification approaches which can scale to millions of SARS-CoV-2 sequences, and clustering methods for preprocessing single-cell cancer data which seems to boost downstream phylogeny inference.

Broader Impacts. The proposal will enhance the knowledge of large-scale evolutionary trends in viruses such as SARS-CoV-2, and of many different tumor types at the resolution of the single cell. The results of this work will be in the form of publications, software and reproducible data analysis pipelines that can be used by other researchers. Since this research is interdisciplinary in nature, it unites academics, researchers and students from biology, data science and mathematics. Such interdisciplinary areas tend to garner participation from women, who are underrepresented in computer science, yet overrepresented in biology. Moreover, Georgia State University is one of the largest minority serving institutions in the U.S., allowing ample opportunity for participation of groups which are historically underrepresented in higher education.