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 evolutionary tree, or *phylogeny*. More than a century later, computational methods for constructing such phylogenies such as maximum parsimony (Fitch, 1971) and neighbour joining appeared. By the 1980s, advances in data collection led to the field of *molecular phylogenetics*, discarding Darwin's purely selectionist theory in favor of a more data-driven neutral theory of evolution (Kimura, 1983). As the field of molecular evolution exploded with data, new computational models and methods based on maximum likelihood (Felsenstein, 1989) and Bayesian approaches (Drummond, 2007) appeared.

Nowadays, with the advent 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 much more diverse. This first issue puts a strain on the *scalability* of existing computational methods for reconstructing phylogenies, while the second issue challenges the generality of current models for representing the data. An example of the first issue 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 methods, based on maximum likelihood, for reconstructing viral phylogenies such as Nextstrain (Hadfield, 2018) can scale to several thousand sequences, while IQTREE-2 (Minh, 2020) can scale to tens of thousands with parallelization. An example of the second issue is single-cell DNA sequencing (scDNA-seq) which allows us to sequence a cancerous tumor at the resolution of a single cell. At this resolution, we widespread evidence of back-mutations (Kuipers, 2017), which 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 are most likely wrong.

The goal of this proposal is to rethink computational phylogenetics from both a *modeling* and *methods* perspective, to work towards developing a general framework which can both scale to the current data, but also to model the increasingly diverse types of data.

Intellectual Merit. TBD

Broader Impacts. TBD