**Computational phylogenetics for algorithms designers**

Phylogenetic trees are used in many biological analyses, including protein structure and function prediction, microbiome anaysis, and the inference of human migrations. Phylogenetic tree construction is commonly posed as a statistical estimation problem where DNA sequences evolve down some unknown tree and the objective is to reconstruct that unknown tree as accurately as possible. Over the last 50 years, many statisticians and probabilists have made great breakthroughs in both models of sequence evolution and analytical methods for estimating phylogenies under these models, and so have transformed the field of computational methods for phylogeny estimation. Indeed, the availability of sophisticated computational methods, fast computers and high performance computing (HPC) platforms, and large sequence datasets enabled through DNA sequencing technologies, has led to the expectation that highly accurate large-scale phylogeny estimation, potentially answering open questions about how life evolved on earth, should be achievable.

Yet, large-scale phylogeny estimation turns out to be much more difficult than expected. First, all the best methods are computationally intensive, and standard techniques do not scale well to large datasets; massive parallelism helps but does not really address the basic challenge inherent in searching an exponential search space. Another issue is that the statistical models of sequence evolution that properly address genomic data are substantially more complex than the ones that model individual loci, and methods to estimate genome-scale phylogenies are (relatively speaking) in their infancy compared to methods for single gene phylogenies. Finally, there is a substantial gap between performance as suggested by mathematical theory (which is used to establish guarantees about methods under statistical models of evolution) and how well the methods actually perform on data – even on data generated under the same statistical models! Indeed, this gap is one of the most interesting things about doing research in computational phylogenetics, because it means that the most impactful research in the area must draw on mathematical theory (especially probability theory and graph theory) as well as on observations from data.

In the last few years, many exciting advances have been made in large-scale phylogeny estimation, by drawing on innovative algorithm design techniques developed in computer science. Some of these new methods are already changing how evolutionary biologists analyze their data, and it is clear that additional new techniques can - and will - enable breakthroughs in biological discovery for the genome-scale datasets that are being assembled around the world.

*"Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation"* aims to train the next generation of algorithm developers so that they can develop these new methods and enable these breakthroughs. A secondary goal is to enable biologists to understand the methods and their statistical guarantees under these models of evolution, so that they can select appropriate methods for their datasets and select appropriate datasets given the available methods. Therefore, although the focus is on communicating mathematical foundations and innovative algorithm design, the book is written to

be accessible to biologists and others with little to no background in computer science, mathematics, and statistics.