**Computational phylogenetics for algorithms designers**

Phylogenetic trees are used in many biological analyses, including protein structure and function prediction, microbiome analysis, and the inference of human migrations. 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.

Computer scientists have brought innovative algorithm design techniques into computational phylogenetics that are dramatically improving the accuracy and scalability of phylogeny estimation. Many of these new methods are now being used by evolutionary biologists to compute multiple sequence alignments, construct species trees and phylogenetic networks from genome-scale datasets, and make biological discoveries. It is clear that computer science 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” is designed to train the next generation of algorithm developers so that they can develop these new methods and enable these breakthroughs. The book is self-contained, and no biology background is needed. Although the focus is on communicating mathematical foundations and innovative algorithm design, much of the material is accessible to biologists and others who are interested in critically evaluating the scientific literature about phylogeny estimation methods in this post-genome era.