Chapter 6

Summary

Our comparative analysis of evolutionary tree reconstruction concludes by recapitulating what factors have made our results possible. We began our work by explaining the pertinence of using evolutionary trees in conjunction with biomolecular sequences such as genes and proteins. The first cursory overview of our context set the framework for viewing the construction of evolutionary trees from homologous sequences as a probabilistic learning problem. In particular, we stated the focal problem of our study as that of learning evolutionary tree topologies from the sample sequences they generate.

The viability of the Markov model for sequence evolution was then explored. Two features of evolution made our recourse to Markov models particularly appropriate: first, that evolution is "memoryless" — inheritance depends solely on the parents and not on the entire history of ancestors —, and secondly, that mutations along different evolutionary branches occur independently from each other at the molecular level. Considering how random taxon sequences form a Markov chain along any path in the tree, we offered an axiomatic definition of phylogeny, noting importantly that the distribution is fully determined by the root sequence distribution and the sequence transition probabilities. A particularly relevant implication of this definition is that the evolutionary tree topology is a function of the joint taxon sequence distribution, making topology recovery from sample sequences at least hypothetically possible. We then set out to investigate the i. i. d. Markov model in which the taxon sequence distribution is a product distribution of identical and independently distributed taxon labels corresponding to sequence characters. An important feature of the i. i. d. Markov model is that the number of parameters defining the phylogeny is finite, and sample sequences

of increasing length convey an increasing amount of information about them. We presented a number of subclasses of the i. i. d. Markov model as natural extensions of commonly used nucleotide substitution rate models, accompanied by novel results on the closedness of corresponding transition matrix classes. We specifically discussed the Jukes-Cantor model, Kimura's two and three parameter models, the Hasegawa-Kishino-Yano model, and the Gojobori-Ishii-Nei model.

We further delved into the problem of topology recovery by discussing the nature of evolutionary distances and similarities, where we defined distance as the logarithm of similarity. By treating distances as functionals of distributions over sequence pairs, we defined several axiomatic properties that evolutionary distances possess, such as additivity along paths and symmetry. We presented the Jukes-Cantor distance, Kimura's three parameter distance, and the paralinear distance, and proved that they exhibit the properties of evolutionary distances. Further, we stated our novel result concerning the uniqueness of evolutionary distances, specifically, that evolutionary distance functions differ by only a constant factor in time-reversible mutation models with constant substitution rates. The additive property of evolutionary distances was particularly important for our purposes, since it implied that topologies could be recovered from distances between sample sequences. This recognition led us to scrutinize methods for estimating evolutionary distances from finite sample sequences. We derived novel upper bounds on the probabilities of large deviations in the cases of Jukes-Cantor distance, Kimura's three parameter distance, and paralinear distance. In each case we showed that the tail probabilities decrease exponentially with the sequence length and the square of the similarities between the sequences involved.

We examined existing algorithmic approaches to evolutionary tree topology reconstruction. We defined computational efficiency as polynomial running time in tree size, and statistical efficiency as successful topology recovery from polynomially long sequences. Both efficiency requirements are essential for recovery of large trees with hundreds or thousands of nodes. We offered a comprehensive overview of maximum-likelihood, character-based, and distance-based algorithms. We noted that exact optimization algorithms that select their output by minimizing a penalty function inevitably address to NP-hard problems, and are not computationally efficient. This difficulty is encountered with maximum-likelihood and character-based methods, as well as with numerical taxonomy- and minimum evolution-related distance-based algorithms. We pointed out the lack of statistical efficiency in the case of

character-based methods due to their statistical inconsistency. We described existing theoretical guarantees for successful topology recovery, specifically, the three-point and four-point conditions. We presented the LogDet metric, which is not an evolutionary distance according to our definition, but satisfies the four-point condition, and can thus serve as a basis for topology recovery with a distance-based algorithm. We analyzed the convergence speed of the LogDet metric estimated from sample sequences. In particular, we derived upper bounds on tail probabilities of the estimation error in a similar form to our upper bounds for empirical evolutionary distances. We used our error bounds to extend existing results on the sample length requirements for distance-based algorithms to recover the topology. We remarked that the sample length bounds for popular distance-base algorithms, including Neighbor-Joining, are generally exponential in tree size.

Given the computational and statistical inadequacies of most existing algorithms, we designed a family of novel distance-based methods satisfying the criteria for statistical and computational efficiency. Our algorithms build evolutionary trees by using triplets of leaves. The algorithms are based on the "Harmonic Greedy Triplets" principle, which originates from our result that in the case of the studied evolutionary distances and the LogDet metric, the error committed in estimating the triplet centers depends on the harmonic average of pairwise similarities between the triplet members. We presented the Basic-HGT and Fast-HGT algorithms, where the former runs in cubic time, and the latter in quadratic time in the number of tree nodes. The algorithms use an input parameter that determines the shortest distance between tree nodes. We presented another quadratic-time algorithm, called HGT-FP, which uses the four-point condition, and eliminates the need for the minimum distance input parameter. We proved that all three algorithms are statistically efficient, and the sample length bounds for the first two match the best asymptotic bounds of other statistically efficient algorithms. In fact, our algorithms are the only known topology recovery algorithms that are provably statistically efficient and run in cubic or quadratic time. Based on simulation experiments, we offered a heuristic way of setting the minimum distance parameter of Fast-HGT by employing the minimum evolution principle. The resulting algorithm, called HGT-ME, runs in $O\left(n^2 \log \ell\right)$ time for a tree with n leaves and sample sequences of length ℓ. We compared the computational and statistical efficiency of our algorithms to the efficiency of many existing methods in simulated experiments. Our goal in the experiments was to evaluate the algorithms' appropriateness for

large-scale phylogeny reconstruction. Running time measurements of existing implementations showed the superior speed of HGT-FP requiring a few seconds to reconstruct trees with thousands of leaves on a desktop computer, in contrast to several minutes, hours, or even days in the case of other algorithms. We compared the success of topology recovery between the HGT-FP, HGT-ME, heuristic parsimony, and several distance-based algorithms, such as Neighbor-Joining. In the experiments we simulated sequence evolution in the Jukes-Cantor model along biologically motivated trees with 135, 500, 1895, and 3135 leaves, with varying mutation probabilities and sample sequence lengths. In summary, we found that heuristic parsimony performs very well, but its slow speed hinders its use with large trees; that Neighbor-Joining is viable in the case of small mutation probabilities but still fails to recover about 1% of the edges from realistic sample lengths, and performs poorly when mutation probabilities are large; and that HGT-FP achieves high success rates when mutation probabilities are large, and fails to recover 5–7 times as many edges as Neighbor-Joining when mutation probabilities are small. Moreover, HGT-FP tends to achieve even higher success rates as the tree size increases. The theoretical results and the simulation experiments show that large-scale phylogeny recovery is feasible with distance-based methods in Markov models of evolution, and our HGT-FP algorithm is particularly useful where other distance-based methods fail. The success of our algorithms is attributable to the greedy selection lying at their core. They do not aim to optimize any explicit penalty function but strive to recover the topology as correctly as possible. As a result, they avoid theoretical and experimental weaknesses of optimization methods.

The path we followed in our dissertation led from molecular sequences and mathematical sequence evolution models, to the design of algorithms with superior efficiency within these models. We hope that in the future we will be able to close this conceptual circle, and that the algorithms will prove useful for molecular evolutionary studies based on biomolecular sequences.