Optimal, Efficient Reconstruction of Root-Unknown Phylogenetic Networks with Constrained Recombination

Dan Gusfield [1][2]

Keywords: Molecular Evolution, Phylogenetic Networks, Ancestral Recombination Graph, Recombination, SNP

1 Introduction

With the growth of genomic data, much of which does not fit ideal evolutionary-tree models, and the increasing appreciation of the genomic role of such phenomena as recombination, recurrent and back mutation, horizontal gene transfer, gene conversion, and mobile genetic elements, there is greater need to understand the algorithmics and combinatorics of phylogenetic networks on which extant sequences were derived. Phylogenetic networks are models of sequence evolution that go beyond trees, allowing biological operations that are not tree-like. One of the most important biological operations is recombination between two sequences.

Hein et al. introduced and studied the *phylogenetic network problem (with recombination)*: Construct a phylogenetic network that derives a given set of binary sequences M, minimizing the number of recombinations used [3, 4, 8, 7, 6, 5]. The minimization criteria is motivated by the general utility of parsimony in biological problems, and because most evolutionary histories are thought to contain a small number of observable recombinations. The assumption that the sequences are binary is motivated today by the importance of SNP data, where each site can take on at most two states (alleles).

No efficient, general algorithm is known for the phylogenetic network problem. Wang et al. [9] introduced a restricted version of the problem: Construct a phylogenetic network when the network is constrained to be a "galled-tree", and the ancestral sequence for the galled-tree is specified in advance. A galled-tree is a phylogenetic network where all cycles (created by recombinations), must be disjoint from each other. Simulations have shown that galled-trees are common when the recombination rate is moderate. The problem of determining whether a set of sequences can be derived on a galled-tree, with a specified ancestral sequence, has an efficient solution [1, 2]. Moreover, when there is a galled-tree for the input, with the specified ancestral sequence, the algorithm produces one that minimizes the number of recombinations over all possible phylogenetic networks with that ancestral sequence. However, the more biologically realistic case is that *no* ancestral sequence is known in advance, and the only previous algorithmic solution for that case takes exponential time.

[1] Department of Computer Science, University of California, Davis, E-mail: gusfield@cs.ucdavis.edu Dan Gusfield

[2] Research partially supported by NSF Grant EIA-0220154. Thanks to C. Langley and S. Eddhu for helpful conversations on this topic.

2 Solution to the Root-Unknown Galled-Tree Problem

We have now developed an efficient solution to the root-unknown galled-tree problem, i.e., in the case when no ancestral sequence is known in advance. For input consisting of n sequences, each of length m, the algorithm runs in $O(nm + n^3)$ time. We show that when there is a galled-tree for the input, the algorithm finds one that minimizes the number of recombinations over all possible phylogenetic networks and over all possible ancestral sequences. This result holds even if multiple-crossover recombinations are allowed.

The main tools that we use to solve the root-unknown galled-tree problem are two graphs representing "incompatibilities" and "conflicts" in M. The conflict graph was used to solve the galled-tree problem when an ancestral sequence is specified in advance [1, 2]. The incompatibility graph is the analogous graph when no ancestral sequence is known. The conflict graphs can be different for different ancestral sequences, so the main difficulty in extending the previous solution to the root-unknown case is that we do not know which conflict graph to use, and there may be an exponential number of them. The main new structural result is that there is always an ancestral sequence A such that its conflict graph is the same as the incompatibility graph. The algorithmic consequence is that even without knowing A, we can determine its conflict graph, and build the network based on that graph, along with other needed modifications to other parts of the previous solution that depend on knowing the ancestral sequence.

References

[1] D. Gusfield, S. Eddhu, and C. Langley. Efficient reconstruction of phylogenetic networks (of SNPs) with constrained recombination. In *Proceedings of 2'nd CSB Bioinformatics Conference*. IEEE Press, 2003.

[2] D. Gusfield, S. Eddhu, and C. Langley. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination, To appear in J. Bioinformatics and Computational Biology. Technical report, UC Davis, Department of Computer Science, 2003.

[3] J. Hein. Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci*, 98:185–200, 1990.

[4] J. Hein. A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.*, 36:396–405, 1993.

[5] R. Hudson and N. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111:147-164, 1985.

[6] S. R. Myers and R. C. Griffiths. Bounds on the minimum number of recombination events in a sample history. *Genetics*, 163:375–394, 2003.

[7] Y. Song and J. Hein. On the minimum number of recombination events in the evolutionary history of DNA sequences. *Journal of Mathematical Biology (to appear)*, 2003.

[8] Y. Song and J. Hein. Parsimonious reconstruction of sequence evolution and haplotype blocks: Finding the minmimum number of recombination events. In *Proc. of 2003 Workshop on Algorithms in Bioinformatics*. Springer-Verlag LNCS, 2003.

[9] L. Wang, K. Zhang, and L. Zhang. Perfect phylogenetic networks with recombination. *Journal of Computational Biology*, 8:69–78, 2001.