Phylogenetic Networks with Constrained and Unconstrained Recombination

Dan Gusfield

Abstract: A phylogenetic network is a generalization of a phylogenetic tree, allowing structural properties that are not tree-like. With the growth of genomic data, much of which does not fit ideal 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.

Wang et al. studied the problem of constructing a phylogenetic network for a set of nn binary sequences derived from a known ancestral sequence, when each site in the sequence can change state at most once in the network, and recombination between sequences is allowed. They showed that the problem of finding a phylogenetic network that minimizes the number of recombination events is NP-hard, but gave a polynomial-time algorithm ($\mathcal{O}(nm + n^4)$-time, for n sequences of length m each) that was intended to determine whether the sequences could be derived on a phylogenetic network in which the recombination cycles are node disjoint. We call such a network a "galled-tree". The paper by Wang et al. is seminal in defining this problem and asserting that it has an efficient solution. Unfortunately, the algorithm given there is incomplete and only constitutes a sufficient, but not a necessary, test for the existence of a galled-tree for the data.

In this talk we do several things. By more deeply analyzing the combinatorial constraints on galled-trees, we obtain an algorithm that runs in $\mathcal{O}(nm + n^3)$-time and is guaranteed to be both a necessary and sufficient test for the existence of a galled-tree for the data. We show how to relax the assumption that we know the ancestral sequence. We show that when there is a galled-tree, the algorithm constructs a "reduced" galled-tree that minimizes the number of mutations occurring on recombination cycles. We prove that the algorithm produces a reduced galled-tree that minimizes the number of recombinations needed for the data, over all possible phylogenetic networks for the data, even if multiple crossovers, or different ancestral sequences are allowed. Hence, when a galled-tree exists, the problem of minimizing the number of recombinations can be solved efficiently. The effect is that the galled-tree is a phylogenetic network that explains the input sequences with the "littlest deviation" from a true tree model. We show that the reduced galled-tree is "nearly-unique", but when it is not unique, the algorithm also obtains a count of the number of galled-trees that exist for the input data, and can create these in linear time for each one, starting from the canonical galled-tree. Finally, we consider phylogenetic networks where the recombination networks are not constrained in any way. We discuss new, efficiently computed lower bounds on the number of recombination events needed.

Joint work with Satish Eddhu, Chuck Langley, and Dean Hickerson