Near-perfect Phylogeny Construction from Genetic Variation Data

Russell Schwartz

The recent accumulation of large amounts of genotype data from human populations has created an unprecedented opportunity for inferring the history of the human species, identifying the forces shaping our genome, and finding links between genotype and phenotype. Making full use of these data will, however, require substantial improvements in algorithms for large-scale phylogenetic inference. Fast algorithms are available for inferring maximally parsimonious mutational histories from genetic variation data under various restrictive assumptions. The "perfect phylogeny" model, which assumes that each genetic variant appears only once within a population's history, has been particularly important to developing phylogenetic algorithms suitable for high-throughput analysis. But such restrictive models are not robust to the variability observed in real data sets. Heuristic methods, which may give solutions far from optimal, thus remain the standard in practice.

In this talk, we will cover recent developments in algorithms for optimal phylogenetic inference based on the search for "near-perfect" phylogenies, those that come close to satisfying the perfect phylogeny assumption. Allowing even small deviations from perfection allows the models to describe a substantially broader group of genotype data. We will examine a recent algorithm proving fixed parameter tractability for the problem of binary parsimony-based inference when the optimal phylogeny differs from a perfect phylogeny by a small number of additional mutations. We will then cover a method for unphased genotype data, a variation of considerable practical importance in which contributions from homologous chromosomes are not separated in the input data. This problem can also be solved in polynomial time when the deviation from perfection is bounded. Finally, we will see some applications of these methods to real variation data.