Species tree estimation from genome scale data

Abstract: Species trees describe the evolutionary relationships between a set of taxa, and are a fundamental tool in various biological analyses. Accurate reconstruction of species trees is complicated by potential discordance between evolutionary histories of various loci in the genome (i.e., “gene trees”) due to biological processes such as Incomplete Lineage Sorting (ILS). In the presence of ILS, statistically consistent “summary methods” have been developed to estimate the species tree given the true gene tree distribution. However, in practice, using these methods on genome-wide sequence data has proved challenging. A basic shortcoming of summary methods is that they assume that the true gene tree distribution is known. However, with genome-wide data, many loci have limited information, resulting in extremely noisy estimates of the gene tree distribution, and consequently inaccurate estimates of the species tree. To address this shortcoming, we have developed a new approach called “statistical binning” that bins genes into sets and uses these bins to estimate the gene tree distribution [1]. Our binning algorithm uses statistical measures of branch support, and is based on a graph-centric optimization problem. We have shown through experimental studies that binning can dramatically increase the accuracy of gene tree distribution, and therefore the accuracy of the species tree. A second challenge was that existing summary methods have reduced accuracy with even a moderately large number of taxa, among other conditions. To enable analyses of larger datasets, we have developed a new summary method, called ASTRAL [2], that can run on datasets with up to 1000 taxa and 1000 genes. We have shown that ASTRAL produces trees that are more accurate than competing summary methods. ASTRAL and statistical binning enabled analyses of two very large datasets, one on 103 plant species [3] and another on 48 bird species [4].