Chapter VI

Efficient and Robust Analysis of Large Phylogenetic Datasets

Sven Rahmann, Bielefeld University, Germany
Tobias Müller, University of Würzburg, Germany
Thomas Dandekar, University of Würzburg, Germany
Matthias Wolf, University of Würzburg, Germany

Abstract

The goal of phylogenetics is to reconstruct ancestral relationships between different taxa, e.g., different species in the tree of life, by means of certain characters, such as genomic sequences. We consider the prominent problem of reconstructing the basal phylogenetic tree topology when several subclades have already been identified or are well known by other means, such as morphological characteristics. Whereas most available tools attempt to estimate a fully resolved tree from scratch, the profile neighbor-joining (PNJ) method focuses directly on the mentioned problem and has proven a robust and efficient method for large-scale datasets, especially when used in an iterative way. We describe an implementation of this idea, the ProfDist software package, which is freely available, and apply the method to estimate the phylogeny of the eukaryotes. Overall, the PNJ approach provides a novel effective way to mine large sequence datasets for relevant phylogenetic information.
Introduction

Basics

Phylogenetic analyses aim to reconstruct ancestral relationships (the phylogeny) between different taxa, e.g., different species in the tree of life, by means of certain morphological or molecular characters. The latter ones are available from the increasing number of sequencing projects. Because of the wealth of data generated in these projects, we need efficient methods to mine the sequence databases for relevant phylogenetic information if we want to reconstruct the most likely phylogeny.

Over the last 50 years, researchers have proposed many methods for phylogenetic tree reconstruction, based on different concepts and models. Each method has its strengths and weaknesses. Distance-based methods, such as Neighbor-joining (Saitou and Nei, 1987) or improved variants thereof, e.g., WEIGHBOR (Bruno, Socci, & Halpern, 2000), BIONJ (Gascuel, 1997), FASTME (Desper & Gascuel, 2002), are relatively fast (the running times generally grows as a cubic polynomial of the number of taxa), but first reduce the information contained in the characters to a matrix of distances. Character-based methods, on the other hand, such as maximum parsimony (Camin & Sokal, 1965; Fitch, 1971), maximum-likelihood (Felsenstein, 1981), or Bayesian methods (e.g., MrBayes; Huelsenbeck & Ronquist, 2001; see Holder & Lewis, 2003, for a review) work directly with character data, but usually require an evaluation of super — exponentially many tree topologies; therefore one reverts to heuristics. There seems to be no universally accepted best method, especially for large datasets.

All of the above-mentioned methods aim to estimate a fully resolved tree from scratch. In some cases, this results in more computational work than needs to be done, or even than the data can robustly support, as usually evidenced by low bootstrap values in the basal branching pattern of the tree. In several analyses, however, we are mainly interested in the basal branching pattern of known or clearly separated and fully supported subclades. In other words, given families of closely related sequences, what is the topology showing the relationships between these families? The problem arises for many large phylogenetic datasets. Müller, Rahmann, Dandekar, and Wolf (2004) recently published a case study on the Chlorophyceae (green algae). Further prominent large phylogenetic datasets in question are the given families of related sequences of, e.g., the Viridiplantae, the Metazoa, the eukaryotes, or even of all the species classified within the tree of life.

As the field of phylogenetics involves both the biological and the mathematical community, the language can sometimes be confusing. For the purpose of this article, we use the terms “subtree” and “subclade” synonymously; each subtree defines a monophyletic group of taxa, i.e., a grouping in which all species share a common ancestor, and all species derived from that common ancestor are included. We usually work with unrooted trees, i.e., connected undirected acyclic graphs, in which every internal node has exactly three neighboring nodes. The taxa are the leaves of the tree.
Related Content

**Evolutionary Algorithm for Identifying Predictive Genes**
[www.igi-global.com/chapter/evolutionary-algorithm-identifying-predictive-genes/53903?camid=4v1a](www.igi-global.com/chapter/evolutionary-algorithm-identifying-predictive-genes/53903?camid=4v1a)

**A New Mechanical Algorithm for Calculating the Amplitude Equation of the Reaction-Diffusion Systems**
[www.igi-global.com/chapter/new-mechanical-algorithm-calculating-amplitude/48377?camid=4v1a](www.igi-global.com/chapter/new-mechanical-algorithm-calculating-amplitude/48377?camid=4v1a)

**Identification Methods of G Protein-Coupled Receptors**
[www.igi-global.com/article/identification-methods-protein-coupled-receptors/73910?camid=4v1a](www.igi-global.com/article/identification-methods-protein-coupled-receptors/73910?camid=4v1a)
High-Accuracy Characterization of Ambulatory Holter Electrocardiogram Events: A Comparative Study Between Walsh-Hadamard Transform, First-Derivative-Based and Intelligent Techniques
www.igi-global.com/article/high-accuracy-characterization-ambulatory-holter/70017?camid=4v1a