The Worst Case Complexity of Maximum Parsimony

Amir Carmel, Noa Musa-Lempel, Dekel Tsur, and Michal Ziv-Ukelson
Department of Computer Science, Ben-Gurion University of the Negev
{karmela,noamu,dekelts,michaluz}@cs.bgu.ac.il

Abstract. One of the core classical problems in computational biology is that of constructing the most parsimonious phylogenetic tree interpreting an input set of sequences from the genomes of evolutionarily related organisms. We re-examine the classical Maximum Parsimony (MP) optimization problem for the general (asymmetric) scoring matrix case, where rooted phylogenies are implied, and analyze the worst case bounds of three approaches to MP: The approach of Cavalli-Sforza and Edwards [5], the approach of Hendy and Penny [12], and a new agglomerative, “bottom-up” approach we present in this paper. We show that the second and third approaches are faster than the first by a factor of \(\Theta(\sqrt{n})\) and \(\Theta(n)\), respectively.

1 Introduction

Phylogenetics is the study of evolutionary relationships among groups of organisms (e.g. species, populations), which are discovered through molecular sequencing data and morphological data matrices. Phylogenies (also called dendograms) are graph-like structures whose topology describes the inferred evolutionary history among a set of biological entities, such as species or DNA sequences. Phylogenies are classically computationally modeled as either rooted or unrooted labeled binary trees, where the input entities are assigned to the leaf vertices. An unrooted phylogeny is an acyclic connected labeled graph in which every vertex has degree of either three or one. Each vertex of degree one has a distinct label. A rooted phylogeny, on the other hand, is similar to an unrooted phylogeny, except that it has one internal vertex of degree two, which is designated as the root. In a rooted phylogeny the edges are directed from the root towards the leaves.

The decision of whether to model phylogenies as rooted versus unrooted trees depends either on the availability of a molecular clock, or on the nucleotide or amino acid substitution scoring matrix representing the evolutionary mutation events. Modeling phylogenies as unrooted trees requires the assumption of symmetric scoring matrices. However, when the symmetry restriction on scoring matrices is removed, the tree rooting becomes meaningful. A simple literature review of current biological research shows that the symmetric scoring matrices, though computationally convenient, do not yield a biologically reliable model [10, 16, 18, 19]. Various recent biological publications apply asymmetric...
scoring matrices to the alignment of genomic sequences, and many papers can nowadays be found on the construction of asymmetric scoring matrices consisting of nucleotides \cite{3,20,21} and amino acids \cite{2,14}. Thus, in this work we do not assume symmetric scoring matrices and therefore construct general, rooted phylogenies.

Phylogenetic trees among a nontrivial number of input sequences are constructed using computational phylogenetics methods. Methods for phylogeny reconstruction can be classified into distance-based versus character-based methods. Given a set of input sequences, a distance-based method, such as UPGMA and neighbor joining (NJ), first computes pairwise distances according to some measure \cite{7}. Then, the actual data is discarded and the fixed distances are used in the derivation of trees. In contrast, in character-based methods (such as Maximum Parsimony and Maximum Likelihood) the inference depends upon models describing the evolutionary changes of characters (e.g. nucleotides or amino acids) that led from an original sequence in some common ancestor to the evolution of the observed input sequences. The great advantage of character-based algorithms for phylogenetic reconstruction is that, given a good multiple alignment of the input sequences, they can exploit the potential phylogenetic inferences with great sensitivity. Their weakness, however, is in their computational intensity. In this paper we focus on the classical Maximum Parsimony character-based phylogenetic approach.

1.1 Phylogenetic Reconstruction Based on Parsimony Maximization

Parsimony Maximization (i.e. preferring the simpler of two otherwise equally adequate theorizations) is one of the classical approaches to computationally reconstruct a phylogeny for a given set of biologically related sequences. When applied to computational phylogenetics, the parsimony maximization approach seeks the phylogenetic tree that supposes the least amount of evolutionary change explaining the observed data \cite{5}. There are two classical problems inferred from phylogenetic parsimony maximization: Small Parsimony (SP) and Maximum Parsimony (MP), explained below.

**Problem 1: Small Parsimony (SP).** The Small Parsimony problem is to compute, for a proposed phylogeny, a reconstruction of events leading to the input data with as few changes as possible over the whole tree. The input to this problem is a multiple alignment of \( n \) input sequences of length \( m \) each, and a topology in the form of a rooted phylogenetic tree over \( n \) leaves, where each leaf is associated with a distinct sequence from the input set. Based on this input, the objective is to compute a labeling of the internal vertices of the input phylogeny which optimizes some predefined scoring scheme.

**Problem 2: Maximum Parsimony (MP).** The Maximum Parsimony problem is to seek, among all possible phylogenies over a given set of leaves, the phylogeny that yields the best SP score. Similarly to SP, the input to this problem is a multiple alignment of \( n \) input sequences of length \( m \) each. However, here the topology is not given. The Maximal Parsimony (MP) problem is NP-Hard \cite{6,9}.