A Multi-objective Evolutionary Approach for Phylogenetic Inference

Waldo Cancino and Alexandre C.B. Delbem

Institute of Mathematics and Computer Science
University of Sao Paulo
Sao Carlos, SP. Brazil 13560–970
{wcancino,acbd}@icmc.usp.br

Abstract. The phylogeny reconstruction problem consists of determining the most accurate tree that represents evolutionary relationships among species. Different criteria have been employed to evaluate possible solutions in order to guide a search algorithm towards the best tree. However, these criteria may lead to distinct phylogenies, which are often conflicting among them. In this context, a multi-objective approach can be useful since it could produce a spectrum of equally optimal trees (Pareto front) according to all criteria. We propose a multi-objective evolutionary algorithm, named PhyloMOEA, which employs the maximum parsimony and likelihood criteria to evaluate solutions. PhyloMOEA was tested using four datasets of nucleotide sequences. This algorithm found, for all datasets, a Pareto front representing a trade-off between the criteria. Moreover, SH-test showed that most of solutions have scores similar to those obtained by phylogenetic programs using one criterion.

Keywords: Phylogenetic Inference, Multi-Objective Optimization, Genetic Algorithms.

1 Introduction

In a recent paper, Handl et al [1] discussed applications of multi-objective optimization in several bioinformatics and computational biology problems. Phylogenetic inference, which searches for the best explanation for evolutionary events from input data, is one of the central problems in this area. It is often modeled as a single objective optimization problem using one criterion for evaluating possible solutions. Moreover, several researches [2,3,4] have shown that the employment of different reconstruction methods can lead to unequal trees for the same input data. Thus, a multi-objective approach, which can search for phylogenies using more than one criterion, can be a relevant contribution since it can produce solutions which are consistent with all employed criteria.

Rokas et al [5] pointed out that there are several sources of incongruence in phylogenetic analysis: optimality criterion employed, data used and evolutionary assumptions about data. Moreover, Poladian and Jermiin [6] suggested that multi-objective optimization can be applied to phylogenetic inference from several conflicting input data. The authors showed that this approach can reveal sources of such conflicts and provide useful information for a robust inference.

S. Obayashi et al. (Eds.): EMO 2007, LNCS 4403, pp. 428–442, 2007.

We propose a multi-objective approach for phylogenetic inference using maximum parsimony [7] and likelihood [8] criteria. The algorithm developed to solve such problem, named PhyloMOEA, is a multi-objective evolutionary algorithm based on NSGA-II model proposed by Deb et al [9]. The output of PhyloMOEA is a solution set representing a trade-off between the criteria considered.

This paper is organized as follows. Section 2 provides relevant background information about phylogenetic inference. Section 3 presents the main concepts of Genetic Algorithms and their application to phylogeny. Section 4 discusses multi-objective optimization problems and shows how Genetic Algorithms can contribute to solve this kind of problems. Section 5 presents the PhyloMOEA algorithm. Section 6 describes the experiments involving four nucleotide datasets and discusses the main results. Finally, Section 7 presents conclusions and proposes future works.

2 Phylogenetic Inference Problem

Phylogenetic analysis investigates evolutionary relationships among species. Sequence data from actual species (nucleotide or aminoacid sequences) are frequently employed for this purpose, although other types of data can be used [10]. Evolutionary relationships can be illustrated as a leaf-labelled tree, named phylogenetic tree. In such tree, external nodes refer to actual species in data, internal nodes refer to hypothetical ancestors and branches represent relations among species. Since sequence data used in phylogenetic analysis are obtained from contemporary species, a phylogenetic tree is a hypothesis (of many possible trees) about the evolutionary events in the history of species.

A phylogenetic tree can be rooted or unrooted. In a rooted tree, there is a special node named root that defines the direction of the evolution, allowing the determination of ancestral relationships among nodes. An unrooted tree shows only the relative positions of nodes without an evolutionary direction. Additionally, tree branches may have an associated length showing genetic distances between connected nodes. Figures 1 and 2 show a rooted and unrooted tree, respectively.

The main goal of the phylogenetic inference is the determination of a tree that best explains the evolutionary events of species under analysis. Swofford et al [11] classified phylogenetic inference methods into two categories: algorithmic and optimality criterion methods. The former follows a sequence of well-defined steps to generate a tree. Clustering methods, like Neighboor Joining [12] are in this category. These algorithms go directly to the final answer without examining manyS alternatives in the search space, quickly producing a phylogenetic tree. Optimally criterion methods include two components: an optimality criterion and a search mechanism. The optimality criterion defines an objective function that scores every possible solution. Using this criterion, the search mechanism should determine the best scored solution in the search space. However, finding the optimal solution requires exhaustive or exact strategies, which are only applicable to small datasets. Since the tree search space increases exponentially