Information Sciences 330 (2016) 293–314

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Performance evaluation of dominance-based and indicator-based multiobjective approaches for phylogenetic inference

Sergio Santander-Jiménez*, Miguel A. Vega-Rodríguez

Department of Computer and Communications Technologies, University of Extremadura, Escuela Politécnica, Campus Universitario s/n, Cáceres 10003, Spain

ARTICLE INFO

Article history:
Received 31 July 2015
Revised 20 September 2015
Accepted 6 October 2015
Available online 22 October 2015

Keywords:
Bioinformatics
Multiobjective optimization
Metaheuristic performance assessment
Phylogenetic reconstruction

ABSTRACT

One of the main research lines in bioinformatics focuses on the optimization of biological processes involving several objective functions. Due to the variety of multiobjective strategies which are available, comparative studies are needed to decide which algorithmic designs lead to improved results. This work tackles the inference of phylogenetic relationships by means of multiobjective metaheuristics. More specifically, we perform the comparative assessment of two lines of multiobjective schemes: dominance-based and indicator-based approaches. On the dominance-based side, we consider two algorithms: Fast Non-Dominated Sorting Genetic Algorithm II and Strength Pareto Evolutionary Algorithm 2. Indicator-based designs are represented by the Indicator-Based Evolutionary Algorithm and a new Indicator-Based Multiobjective Bat Algorithm. The experimental evaluation of these methods is conducted over six real biological datasets, making comparisons with multiple state-of-the-art phylogenetic tools. Our experimentation verifies the significant performance achieved when combining indicator-based approaches and swarm intelligence. Particularly, different multiobjective metrics (hypervolume, set coverage, and spacing) and biological testing procedures highlight the promising results reported by this kind of algorithmic designs.

1. Introduction

Throughout the years, bioinformaticians have focused their research efforts on the design of computational approaches to deal with NP-hard biological problems. The modelling of such processes as optimization problems seeks to find the most satisfying solution in accordance with an optimality criterion which assesses its biological relevance. Thanks to the advances in algorithmic development, more realistic assumptions have been incorporated into these models, leading to the need to tackle problems which involve multiple criteria simultaneously. This reason explains the interest on applying multiobjective optimization techniques to bioinformatics [21]. Given a decision space S and an objective space $Z = \Re^n$, a multiobjective optimization problem (MOP) consists of finding those solutions $s = (s_1, s_2, \ldots, s_k) \in S$ (defined by k decision variables) which optimize n objective functions $\vec{f}(s) = (f_1(s), f_2(s), \ldots, f_n(s)) \in Z$ [10]. Due to the infeasible nature of finding the Pareto-optimal set in most real-world problems, bioinspired and evolutionary multiobjective methods have been defined in order to obtain good Pareto set approximations in reasonable times.

* Corresponding author. fax.: +34 927 257000x57574
E-mail addresses: sesaji@unex.es (S. Santander-Jiménez), mavega@unex.es (M.A. Vega-Rodríguez).

http://dx.doi.org/10.1016/j.ins.2015.10.021
0020-0255/

Phylogenetic reconstruction represents one of the most relevant optimization problems in bioinformatics [26]. Given a set of molecular sequences from different organisms, phylogenetic procedures aim to explain the features observed in a set of species by inferring their evolutionary history. Such knowledge gives support to a wide number of scientific fields, including epidemiological dynamics, comparative genomics, and cancer research. One of the main reasons behind the formulation of this problem as a MOP is given by the need to address incongruence issues related to the choice of the optimality criterion. Several reports [4,28,41] have given account of the inference of conflicting phylogenetic relationships under different optimality criteria (such as parsimony and likelihood [26]). The use of multiobjective optimization has been shown to tackle the problem successfully, providing evolutionary histories which solve the conflicts caused by single-criterion methods [9,37,38].

An additional issue to be addressed is the high computational complexity of the problem, mostly motivated by the explosively growth of the phylogenetic tree search space with the number of species, along with rising evaluation times which depend on sequence length. Hence, exhaustive searches cannot be applied, leading to an increasing interest in developing bioinspired approaches for phylogenetics. In this work, we aim to carry out the comparative analysis of different multiobjective approaches to address this complex biological problem. In this sense, research on the design of multiobjective metaheuristics has evolved throughout the years, defining different algorithmic proposals in accordance with the way they carry out the search for Pareto sets. On the one hand, dominance-based approaches define their design on the basis of the concept of Pareto dominance, which allows the ranking of the solutions managed by the algorithm at each generation. This mechanism is usually complemented by using density estimation measurements to refine the search. Two examples of dominance-based proposals are the Strength Pareto Evolutionary Algorithm 2 (SPEA2) [52] and the Non-Dominated Sorting Genetic Algorithm II (NSGA-II) [16].

On the other hand, an increasingly popular design trend is the use of quality indicators [53] to drive the search. Let Ω be a set containing all the possible Pareto set approximations. A quality indicator I refers (in its unary form) to a function $f{:}\Omega \rightarrow \Re$ which measures the quality of the outcome generated by an algorithm in terms of convergence to the Pareto-optimal front and/or diversity. Those multiobjective approaches which integrate such metrics for fitness assignment purposes are known as indicator-based algorithms. The definition of the optimization goal in terms of quality indicators was suggested in [51], where a general framework for indicator-based optimization was proposed: the Indicator-Based Evolutionary Algorithm (IBEA).

In previous researches, we showed the relevance of applying dominance-based designs such as NSGA-II [38] and nature-inspired search techniques [37,39] to phylogenetics. This paper aims to go a step further in the development of multiobjective solutions to tackle this kind of high-complexity optimization problems. To this end, we undertake a comparative study among different dominance-based and indicator-based multiobjective approaches for inferring phylogenies attending to the parsimony and likelihood principles. Four different metaheuristics are considered: NSGA-II, SPEA2, IBEA, and the Indicator-Based Multiobjective Bat Algorithm (IMOBA), a new multiobjective proposal based on the Bat Algorithm [49]. This performance analysis will be conducted by experimentation over six real nucleotide datasets, using unary and binary multiobjective metrics to decide which approach shows the best overall behaviour. In addition, biological quality will be examined by making comparisons with different state-of-the-art methods.

This paper is organized as follows. In the next section, we give account of related researches published in the literature. The basis of this problem and its formulation as a MOP are summarized in Section 3. Section 4 describes the general features of the multiobjective proposals for phylogenetics reviewed in this paper. In Section 5, we report experimental results and comparisons attending to multiobjective and biological quality. Section 6 is focused on the discussion of the observed results. Finally, the conclusions of this study and future work lines can be found in Section 7.

2. Related work

Throughout the years, the growing availability of genetic data has motivated the development of different algorithmic strategies to satisfy biological processing needs. Phylogenetic procedures must deal with the problem of exploring very large tree search spaces which grow exponentially with the number of species, representing a challenging problem from a computational perspective. The first attempt to apply evolutionary computation to phylogenetics was reported in 1995 by Matsuda [29], who published a genetic algorithm for maximum likelihood inference from amino acid sequences. Later on, Lewis proposed GAML [27], a genetic algorithm which significantly reduced the computational effort needed to carry out maximum likelihood analyses on nucleotide data. In [32], Moilanen introduced PARSIGAL, a hybrid proposal for maximum parsimony which combined evolutionary algorithms and branch-swapping local search procedures. A genetic algorithm with self-adaptive control parameters for inferring likelihood-based phylogenies was proposed by Skourikhine in [45]. Other designs like GAPHYL considered the management of parallel subpopulations [12] to improve solution quality in parsimony-based analysis, outperforming the widely-used PHYLIP tools [17]. On the other hand, different developments focused on the exploitation of hardware resources to accelerate inference times. We can highlight the studies undertaken by Katoh et al. [25] and Brauer et al. [2], who proposed parallel genetic algorithms to conduct computationally demanding analyses.

With the publication of high-complexity datasets, research efforts focused on discussing new techniques to achieve an efficient processing of the tree search space. For example, Cotta and Moscato reported in [13] direct and indirect tree encoding strategies for performing analyses under a distance-based optimality criterion, getting meaningful results at a low computational cost. Poladian studied in [34] the behaviour of a genetic algorithm for maximum likelihood with matrix-shaped representation, using the neighbour-joining method for genotype-phenotype mapping purposes. The proposal showed a significant performance in comparison with other heuristic-based methods like DNAML [17]. In [54], Zwickl proposed a genetic algorithm for rapid likelihood inference, whose design aimed to refine the search at topological, branch length and parameter setting levels. In 2010,