A Comparative Study on Distance Methods Applied to a Multiobjective Firefly Algorithm for Phylogenetic Inference

Sergio Santander-Jiménez
Department of Technologies of Computers and Communications
University of Extremadura (Spain)
sesaji@unex.es

Miguel A. Vega-Rodríguez
Department of Technologies of Computers and Communications
University of Extremadura (Spain)
mavega@unex.es

ABSTRACT

Throughout the years, researchers have reported a wide variety of proposals to infer evolutionary histories from biological data. Recent studies suggested the use of matrices of genetic distances to represent phylogenetic topologies in population-based metaheuristics. A key question that must be addressed is the choice of a particular method to build phylogenies from evolutionary distances. In addition to this, there is a growing need to overcome the problems that arise when different optimality criteria describe conflicting hypotheses about the evolution of the input species. In this paper, we tackle the phylogenetic inference problem by using a multiobjective algorithm with matrix representation inspired by the bioluminescence of fireflies. Our main goal is to study the behaviour of several clustering and neighbor-joining methods applied to infer phylogenies from the distance matrices processed by our algorithm. Experimental results on four real nucleotide data sets point out the advantages and disadvantages of each proposal, in terms of multiobjective performance and processing times.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics; I.2.8 [**Problem Solving, Control Methods, and Search**]: Heuristic methods; G.1.6 [**Optimization**]: Global optimization

General Terms

Algorithms

Keywords

Swarm Intelligence; Multiobjective Optimization; Firefly Algorithm; Phylogenetic Inference; Distance Methods

1. INTRODUCTION

The understanding of the evolutionary events that gave rise to modern species in Nature represents one of the most outstanding research topics in the field of Bioinformatics. A phylogeny is a tree-shaped structure that describes ancestral relationships among species. Phylogenetic analyses contribute interesting knowledge not only in Evolutionary Biology, but also in a wide variety of scientific fields, such as chemistry, medicine, paleontology, and ecology.

Throughout the years, there has been a growing interest in developing new methodologies that formulate phylogenetic inference as an optimization problem. The main goal is the reconstruction of phylogenies according to a specific principle or optimality criterion. In this sense, we must bear in mind several key issues. Firstly, we need to define new approaches to overcome the problems that arise when applying classical methods. Modern biological data sets cannot be analyzed by using exhaustive searches, due to the exponential growth of the number of topologies in the tree search space in accordance with the number of species under review [9]. Novel developments based on Bioinspired Computing have been applied successfully to resolve this problem [11].

In second place, a wide variety of optimality criteria and methods for inferring phylogenies can be found in the literature [17]. We can distinguish two main groups: character-based methods, and distance-based methods. The first procedures aim to infer evolutionary relationships among organisms by processing directly their molecular sequences. In DNA analyses, these sequences are represented as strings of characters according to the alphabet $\Sigma = \{A, C, G, T\}$. Some examples of character-based approaches are maximum likelihood [8] and maximum parsimony [10].

On the other hand, distance-based methodologies take as input a matrix of genetic distances calculated from an estimation of the number of substitution events in molecular chains, according to an evolutionary model that defines the nature and the occurrence probability of such events. Examples of approaches based on distances are clustering methods [26] and neighbor-joining methods [23]. Distance matrices can also be applied as a useful way to represent individuals in population-based algorithms for inferring phylogenies attending to character-based criteria [21].

When choosing a particular optimality criterion, we must take into account the principles that define each methodology. For example, while maximum likelihood seeks to find the most likely genealogical relationships in accordance with statistical measurements, maximum parsimony aims to find the evolutionary tree that minimizes the amount of mutation events throughout the tree. The decision on which method should be used is not a trivial issue, as diverse optimality criteria can give as a result the inference of conflicting re-

GECCO'13 Companion, July 6–10, 2013, Amsterdam, The Netherlands.

lationships from the same biological data [30]. This is one of the reasons why phylogenetic inference can be formulated as a Multiobjective Optimization Problem (MOP) [4]. By means of multiobjective optimization, we aim to reconstruct evolutionary histories that suppose an agreement between different theories about the way species evolve in Nature.

In this paper, we propose a comparative study on distance methods applied to a Multiobjective Firefly Algorithm (MOFA) for inferring phylogenetic trees according to maximum parsimony and maximum likelihood [24]. Our main goal is to assess the performance achieved by the proposal when five tree-building methods are considered to infer phylogenetic topologies from their corresponding matrix representations: neighbor-joining (NJ), BIONJ, unweighted and weighted-pair group methods with arithmetic means (UPGMA and WPGMA), and complete-linkage clustering. Experimental results on four real nucleotide data sets will be evaluated from a multiobjective view by using the hypervolume metrics [29]. In addition to this, we will assess the relevance of our proposal by comparing it with other authors' multiobjective approaches and single-criterion biological methods.

This paper is organized as follows. In Section 2, we summarize some of the most relevant approaches for inferring phylogenies proposed in the literature. Section 3 introduces the basis of phylogenetic inference. In Section 4, we detail the different tree-building methods which are the scope of this research. In Section 5, we explain our swarm intelligence algorithmic design. Section 6 summarizes our experimental methodology and shows experimental results. Finally, Section 7 provides conclusions and proposes future work.

2. RELATED WORK

The enormous tree search space that must be considered even with a low number of species motivate the NP-hard nature of phylogenetic inference [9]. In recent years, new bioinspired and evolutionary proposals have arisen as an answer to the needs of biologists. The design of efficient procedures and the growing availability of biological sequences bring closer the achievement of the ultimate goal in Phylogenetics: the reconstruction of the Tree of Life. In this section, we give account of several well-known bioinspired approaches to phylogenetic inference proposed by other authors.

In 1990's, Matsuda [20] and Lewis [19] laid the foundations of Evolutionary Computation for phylogenetic reconstruction under the maximum likelihood criterion. Matsuda proposed in 1995 the first genetic algorithm for protein phylogenetic inference from amino acid data. In 1998, Lewis developed a genetic algorithm which reduced the computational times required by traditional heuristic-based methods to analyze nucleotide data sets.

Later on, several authors, such as Skourikhine [25] and Lemmon and Milinkovitch [18], followed this line of research, proposing auto-adaptative and multipopulation genetic algorithms to maximum likelihood. Phylogenetic analyses under the maximum parsimony criterion were also tackled by using evolutionary proposals, such as Congdon's proposal: GAPHYL [5]. The introduction of memetic algorithms to hierarchical clustering reconstruction from distance matrices was suggested by Cotta and Moscato in [7], showing significant results in comparison with traditional branch-and-bound techniques and other evolutionary proposals.

One key question that must be considered when designing bioinspired approaches to Phylogenetics is how to represent phylogenetic trees. Gottlieb et al. showed in [15] that the use of Prüfer sequences, a classical indirect representation, leads to poor performances. In [6], Cotta and Moscato discussed a variety of direct and indirect representations applied to reconstruct distance-based phylogenies. In 2005, Poladian achieved meaningful results under the maximum likelihood criterion by using an indirect representation based on distance matrices and neighbor-joining as a genotype-phenotype mapping method [21]. Poladian's work proposed new genetic operators to act directly on distance matrices, which give as a result topological transformations in the shape of the inferred trees.

As new developments allow biologists to carry out efficient analyses on modern biological data, novel trends of research aim to address incongruences in Phylogenetics. By modelling phylogenetic inference as a MOP, researchers' efforts focus on solving two key issues: firstly, the reconstruction of evolutionary histories from sources of data with conflicting information about ancestral relationships, and secondly, the inference of trade-off phylogenies that optimize simultaneously two or more optimality criteria. Poladian and Jermiin addressed the first problem and developed the first multiobjective evolutionary algorithm proposed to Phylogenetics [22]. Afterwards, several authors applied multiobjective metaheuristics following the second research line. Coelho et al. [3] designed an immune-inspired algorithm to reconstruct phylogenetic trees by minimizing the mean-squared error and the minimal evolution criteria. Cancino and Delbem proposed PhyloMOEA [1], a multiobjective genetic algorithm for phylogenetic reconstruction attending to maximum parsimony and maximum likelihood, achieving promising results on four real nucleotide data sets.

We address in this work multiobjective phylogenetic inference according to multiple optimality criteria, in particular, maximum parsimony and maximum likelihood. For this purpose, we propose a study on the performance achieved by a swarm intelligence proposal with matrix-based individual representation when different methods are considered to reconstruct phylogenetic topologies from distance matrices.

3. PHYLOGENETIC INFERENCE

Phylogenetic methods analyze sequences of biological data that characterize a set of N species, with the aim of inferring their ancestral evolutionary relationships. Each sequence is composed by S sites or characters, which contain genetic information, such as nucleotides, amino acids, and morphological data [17]. While character-based methods operate directly over biological data, distance methods consider a matrix of genetic distances generated from these sequences. By processing such molecular or numerical data, we can generate a tree-shaped structure $T = (V, E)$ that represents a hypothesis about the evolutionary events that gave as a result the species characterized by the input data.

In a phylogenetic tree T, the results of the evolutionary history, this is, the input species, are located in the leaves, as terminal nodes in V. Internal nodes in V represent hypothetical ancestors, which are connected to their descendants by branches in E. A branch can be associated to a float type value known as branch length, which defines the evolutionary distance between two related species. This distance can be measured in terms of evolutionary times or mutation events which motivated changes in genetic sequences.

Optimality criteria methods seek to describe an evolution-

ary history according to an objective function that assess the quality of the inferred phylogeny. In this work, we will perform multiobjective phylogenetic analyses according to two criteria: parsimony and likelihood.

3.1 Maximum Parsimony

The evolution of genetic sequences depends on the occurrence of mutation events which motivate morphological changes. A maximum parsimony approach aims to reconstruct those evolutionary histories that *minimize* the number of mutation occurrences to explain the observed data. According to Ockham's razor, parsimony approaches consider that the simplest evolutionary hypothesis should always be preferred. When computing parsimony trees, ancestral sequences must be assigned from the knowledge provided by the input data. Fitch's proposal [10] can be applied to compute a set of ancestral assignments that minimizes the amount of molecular changes throughout the tree.

Given a phylogenetic tree $T = (V, E)$ inferred from a dataset containing N sequences of S sites, its parsimony score $P(T)$ is given by the following equation [13]:

$$P(T) = \sum_{i=1}^{S} \sum_{(a,b)\in E} C(a_i, b_i) \quad (1)$$

where $(a, b) \in E$ represents a genealogical relationship between two nodes a and b, a_i and b_i the values at the ith site of the molecular sequences which characterize a and b, and $C(a_i, b_i)$ the cost of evolving from a_i to b_i. The most parsimonious hypothesis will be that phylogenetic topology which minimizes $P(T)$.

3.2 Maximum Likelihood

Likelihood is a statistical measurement that can be used to conduct phylogenetic analyses [8]. The main goal is the inference of the most likely evolutionary history from the sequences observed in the input data. Likelihood is one of the most widely-used criteria in Phylogenetics, as it was defined in the basis of statistical reliability. However, such reliability implies complex computational requirements due to the high number of parameters involved in likelihood computations, including mathematical models that describe the probability of observing mutation events at molecular level. These evolutionary models must consider several factors related to the way sequences evolve (such as transition/transversion ratios and among-site rate variations) [17] in order to avoid misleading results due to wrong assumptions.

Let D be the observed data, a set of N molecular sequences which contains S sites per sequence, m an evolutionary model, and $T = (V, E)$ a phylogenetic tree inferred from D. We define likelihood as the conditional probability of D given a hypothesis modelled by T and m [9]:

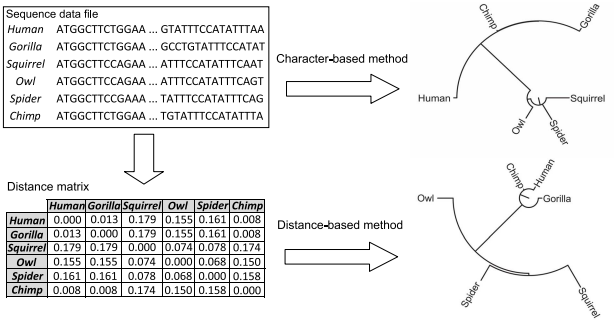
$$L[D, T, m] = \Pr[D|T, m] = \prod_{i=1}^{S} \prod_{j=1}^{E} (r_i t_j)^{n_{ij}} \quad (2)$$

where r_i is the mutation probability for the site i, t_j the evolutionary time between the nodes related by $j \in E$, and n_{ij} the number of changes observed between the nodes connected by j at the ith site. The phylogenetic tree that maximizes the likelihood function will be considered as the most likely hypothesis under the assumptions given by m. In this work, likelihood values will be computed according to the $HKY85 + \Gamma$ evolutionary model [9].

4. DISTANCE-BASED METHODS

Distance approaches [2] are among the most popular methods in Phylogenetics. Distances describe an estimation of the rate of substitutions between molecular sequences, according to the assumptions given by an evolutionary model [17]. As these methods operate over genetic distances, we can distinguish two main tasks to be performed: the computation of distance matrices from molecular data, and the reconstruction of phylogenies from distance matrices.

Given N sequences that characterize the input species, distance-based methods compute in first place a matrix data structure M composed by N rows and N columns, where each entry $M[i, j]$ contains the genetic distance between the sequences that characterize the species (or groups of species) i and j. In a second step, these methods build the phylogeny by grouping iteratively the pair of species in M which optimize a specific criterion. Figure 1 shows an example of this methodology, in comparison with a character-based one.



	Human	Gorilla	Squirrel	Owl	Spider	Chimp
Human	0.000	0.013	0.179	0.155	0.161	0.008
Gorilla	0.013	0.000	0.179	0.155	0.161	0.008
Squirrel	0.179	0.179	0.000	0.074	0.078	0.174
Owl	0.155	0.155	0.074	0.000	0.068	0.150
Spider	0.161	0.161	0.078	0.068	0.000	0.158
Chimp	0.008	0.008	0.174	0.150	0.158	0.000

Figure 1: Character-based and distance-based phylogenetic reconstruction

The computation of genetic distances can also be carried out by considering a phylogenetic topology with branch length values. In this way, we can establish a correspondence between distance matrices and phylogenetic topologies. In this study, we will focus on two tree-building strategies: clustering analysis and neighbor-joining.

4.1 Clustering Methods

Clustering analysis is a traditional approach to build phylogenetic topologies from distances. These methods are defined on the basis of the molecular clock assumption, which considers that the rates of molecular changes in sequences show a clocklike behaviour [9]. The results of applying these techniques are ultrametric rooted trees, which describe equidistant paths from the root node to any terminal node.

Given a distance matrix M, a clustering method proceeds by selecting the entries corresponding to the species or groups of species i,j that minimize the genetic distance value. A partial phylogeny or *cluster of species* is generated by connecting i and j to a new node n_{ij} which represents their common ancestor. Branch length values are then computed and M is updated with the distances to the new cluster, which replace the entries corresponding to i and j. The algorithm follows these steps until M has been completely processed, returning the inferred phylogeny. Algorithm 1 shows the pseudocode for a generic clustering procedure.

Algorithm 1 Generic Clustering Algorithm

1: **repeat**
2: $i, j \leftarrow$ Select the indices i, j which minimize $M[i, j]$
3: $n_{ij} \leftarrow$ Create a new parent node and connect i, j to it
4: */*Assign branch lengths l_i and l_j from i and j to n_{ij} */*
5: $n_{ij}.l_i \leftarrow M[i, j]/2$
6: $n_{ij}.l_j \leftarrow M[i, j]/2$
7: */*Compute distances from n_{ij} to the remaining groups*/*
8: **for** k = 1 to N **do**
9: $D[k] \leftarrow$ Compute distances (M, n_{ij}, k) /* $k \neq i, j$ */
10: **end for**
11: $M, N \leftarrow$ Update matrix (n_{ij}, D)
12: $T \leftarrow$ Add new node (n_{ij})
13: **until** There are no groups to be proccessed

Most clustering methods follow different principles to compute the distances to the new cluster (line 9 in Algorithm 1). We will focus on three clustering methods: UPGMA, WPGMA, and complete-link clustering.

UPGMA [27] assigns the updated values in M by averaging the distances according to the number of elements in the new cluster. Let r be the number of species in a partial phylogeny. The distance from the cluster generated by grouping i and j to each other group k is computed as follows:

$$D[k] = \left(\frac{i.r}{i.r + j.r}\right) M[i, k] + \left(\frac{j.r}{i.r + j.r}\right) M[j, k] \quad (3)$$

On the other hand, WPGMA [26] considers that all the elements included in the new cluster have the same weight in the computation of the distances:

$$D[k] = \frac{M[i, k] + M[j, k]}{2} \quad (4)$$

Finally, in complete-link clustering [16], $D[k]$ refers to the largest distance from the items in the new cluster to k:

$$D[k] = \max(M[i, k], M[j, k]) \quad (5)$$

4.2 Neighbor-joining

The neighbor-joining (NJ) method was originally proposed by Saitou and Nei [23]. This approach differs from clustering analysis in the fact that NJ does not assume a molecular clock. As rates of molecular changes can evolve in different ways among lineages, NJ addresses the analysis of complex data sets by taking into account more realistic assumptions.

Algorithm 2 summarizes the main tasks in this algorithm. Divergence values for each species (line 4 in Algorithm 2) are considered when computing the length of the branches that connect i and j to the new parent node n_{ij} (lines 9 and 10). The assignment of new distances also differs from clustering methods, defining $D[k]$ as $(M[i, k] + M[j, k] - M[i, j])/2$.

In the literature we can find new algorithmic designs which try to modify this scheme. A well-known approach based on NJ is BIONJ, reported by Gascuel [12]. This algorithm was proposed with the aim of improving NJ in terms of statistical error [9]. For this purpose, BIONJ introduces a model of variances and covariances of evolutionary distances. The main goal is to compute a factor λ which minimizes the sampling variances of distances in each iteration of the algorithm. λ is applied to compute the distances from the partial phylogeny to each other group k as follows:

$$D[k] = \lambda M[i, k] + (1 - \lambda) M[j, k] - \lambda n_{ij}.l_i - (1 - \lambda) n_{ij}.l_j \quad (6)$$

Algorithm 2 Neighbor-joining Algorithm

1: **repeat**
2: */*For each row i in M compute the divergence value u[i]*/*
3: **for** i = 1 to N **do**
4: $u[i] \leftarrow \sum_{j(j \neq i)}^{N} M[i, j]/(N - 2)$
5: **end for**
6: $i, j \leftarrow$ Select the indices i, j which minimize $M[i, j] - u[i] - u[j]$
7: $n_{ij} \leftarrow$ Create a new parent node and join i, j to it
8: */*Assign branch lengths l_i and l_j from i and j to n_{ij} */*
9: $n_{ij}.l_i \leftarrow \frac{1}{2} M[i, j] + \frac{1}{2}(u[i] - u[j])$
10: $n_{ij}.l_j \leftarrow \frac{1}{2} M[i, j] + \frac{1}{2}(u[j] - u[i])$
11: */*Compute distances from n_{ij} to the remaining groups*/*
12: **for** k = 1 to N **do**
13: $D[k] \leftarrow (M[i, k] + M[j, k] - M[i, j])/2$ /* $k \neq i, j$ */
14: **end for**
15: $M, N \leftarrow$ Update matrix (n_{ij}, D)
16: $T \leftarrow$ Add new node (n_{ij})
17: **until** There are no more than two groups to be proccessed
18: */*Connect final nodes u, v by a branch with length $M[u, v]$*/*
19: $T \leftarrow$ Add new branch (u, v, $M[u, v]$)

By comparing these distance methods, we try to discuss which approach can lead our proposal to achieve improved performances from a multiobjective point of view.

5. A BIOINSPIRED APPROACH FOR INFERRING PHYLOGENIES

The Multiobjective Firefly Algorithm (MO-FA) is a multiobjective adaptation of a novel bioinspired algorithm proposed by Yang in 2010, Firefly Algorithm (FA) [31]. This proposal belongs to a family of metaheuristics inspired by the swarm intelligence of social insects and other organisms in Nature. A communication system based on brightness and attractiveness governs the behaviour of fireflies. By means of flashing lights, fireflies are able to attract partners to their position. This attraction system depends on three factors: the light intensity, the distance between fireflies, and the degree of light absorption by the environment.

Fireflies with the brightest flashing light patterns will be preferred by their potential partners. In an algorithmic context, this collective behaviour is modelled by identifying the light intensity with the quality of the solution associated to a firefly. In this way, the firefly population will seek for high-quality solutions, conducting the exploration of the search space according to the information provided by the swarm. In our multiobjective proposal, the *dominance* concept [4] is applied to assess the quality of solutions.

In order to adapt this algorithm to phylogenetic inference, we must define a proper individual representation. In this work, phylogenetic trees will be represented by their corresponding distance matrices, generated according to the evolutionary distances given by branch length values. The algorithm will operate over these matrices and phylogenetic topologies will be inferred by using the distance-based methods explained in Section 4.

MO-FA takes as input the following parameters, defined according to the elements that affect the behaviour of fireflies: *dataset* (data to be analyzed), *swarmSize* (population size), *numGenerations* (maximum number of generations), β_0 (attractiveness factor), γ (absorption coefficient), and α (randomization factor). The output will be a set of *Pareto* solutions according to the parsimony and likelihood principles. Algorithm 3 summarizes the main tasks in MO-FA.

Firstly, fireflies are initialized by selecting phylogenetic

Algorithm 3 MO-FA Scheme

```
Initialize the population (X, dataset, swarmSize)
for numIter = 1 to numGenerations do
    for r = 1 to swarmSize do
        for s = 1 to swarmSize do
            if X_s ≻ X_r then
                Move X_r towards X_s (X_r.M, X_s.M, β0, γ, α)
            end if
        end for
        Apply a distance method to infer the tree X_r.T (X_r.M)
        Evaluate solution (X_r.T)
    end for
    Update Pareto set with the most promising solutions (X)
end for
Return Pareto set
```

topologies from a repository of 1000 phylogenies, 500 generated by maximum parsimony techniques, and the remaining 500 by maximum likelihood. These starter trees are evaluated and their distances matrices are computed afterwards.

Each firefly associated to a dominated solution will be updated with the knowledge provided by the most promising phylogenies generated by the algorithm. Let X_r be a firefly dominated by X_s, and $X_r.M$, $X_s.M$ the NxN distance matrices related to X_r and X_s, respectively. We compute the overall distance δ_{rs} that separates X_r from X_s as follows:

$$\delta_{rs} = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{i}(X_r.M[i,j] - X_s.M[i,j])^2} \quad (7)$$

Once δ_{rs} has been calculated, we update each entry in $X_r.M$ by using a movement formula which takes into account the attractiveness factor β_0, the absorption coefficient γ, and the randomization factor α. This last parameter introduces some randomness with the aim of allowing the algorithm to perform searches on undiscovered regions of the search space. Equation 8 describes this updating step.

$$\begin{aligned} X_r.M[i,j] =& X_r.M[i,j] + \beta_0 e^{-\gamma\delta_{rs}^2}(X_s.M[i,j] - X_r.M[i,j]) \\ &+ \alpha(rand[0,1] - \frac{1}{2}) \end{aligned} \quad (8)$$

After the distance matrix has been modified, the resulting phylogenetic topology is inferred by using one of the tree-building methods we consider in this study. When we use for this purpose clustering analysis, a gradient optimization step is carried out to correct branch length values on non-ultrametric data sets. Finally, parsimony and likelihood scores are computed for the inferred phylogeny.

These steps are repeated until all dominated fireflies have been processed. In the final steps of a generation, the set of most promising Pareto solutions is updated, and the extreme points in this Pareto set are optimized by using a topological search methodology [13]. A more detailed explanation and further details about MO-FA can be found in [24]

6. EXPERIMENTAL RESULTS

In this section we explain the experimental methodology we have followed to assess the performance achieved by MO-FA when different tree-building methods are considered. We have configured the input parameters of the algorithm in accordance with our previous work [24], where different combinations of β_0, γ, and α were studied to find an optimal configuration. The final values for each parameter are the following: swarmSize = 100, numGenerations = 100, β_0 = 1, $\gamma = 0.5$, and $\alpha = 0.05$.

Table 1: Experimental methodology: data sets

Dataset	Sequences	Sites per sequence	Description
rbcL_55	55	1314	rbcL gene from green plants
mtDNA_186	186	16608	Human mitochondrial DNA
RDPII_218	218	4182	Prokaryotic RNA
ZILLA_500	500	759	rbcL plastid gene

In this study, we have performed phylogenetic analyses under the $HKY85 + \Gamma$ evolutionary model on four real nucleotide data sets from the literature [1], given by Table 1. 30 independent runs of the algorithm were carried out for each dataset and distance method. Phylogenetic results have been evaluated from a multiobjective perspective by applying the well-known hypervolume metrics [29]. Table 2 gives account of the reference points used to compute hypervolume. The processing times required by each configuration have also been captured (using 16 OpenMP threads), in order to make possible a comparison in terms of quality of solutions and computational complexity.

Table 2: Reference points for hypervolume

	Ideal reference point		Nadir reference point	
Dataset	Parsimony	Likelihood	Parsimony	Likelihood
rbcL_55	4774	-21569.69	5279	-23551.42
mtDNA_186	2376	-39272.20	2656	-43923.99
RDPII_218	40658	-132739.90	45841	-147224.59
ZILLA_500	15893	-79798.03	17588	-87876.39

Experimental results are presented in Table 3. For each dataset and distance method, we show the parsimony and likelihood scores corresponding to the best parsimony tree (columns 2-3) and the best likelihood tree (columns 4-5) from the execution which achieved the closest value to the overall mean hypervolume, given by columns 6-7. The average processing times required for each proposal are given by column 8. Figure 2 shows the corresponding Pareto fronts generated for each dataset, according to the tree-building method which was applied to perform the analysis.

Table 3: Experimental results

rbcL_55

	Best parsimony tree		Best likelihood tree		Hypervolume metrics		Execution
Methods	Pars.	Like.	Pars.	Like.	Mean	Std. Dev.	Times (s)
UPGMA	4880	-21891.084	4882	-21885.359	66.488%	0.352	287.613
WPGMA	4874	-22085.094	4881	-21893.387	66.916%	0.481	269.203
Complete	4874	-22456.859	4885	-21898.369	66.491%	0.619	**244.188**
NJ	4874	-21857.617	4891	-21818.107	**70.050%**	**0.067**	395.729
BIONJ	4874	-21852.010	4891	-21819.188	70.039%	0.070	399.001

mtDNA_186

	Best parsimony tree		Best likelihood tree		Hypervolume metrics		Execution
Methods	Pars.	Like.	Pars.	Like.	Mean	Std. Dev.	Times (s)
UPGMA	2451	-39985.988	2453	-39984.811	61.416%	1.307	3476.906
WPGMA	2453	-40054.581	2467	-39992.076	61.215%	0.827	3473.117
Complete	2451	-40014.640	2454	-39979.645	62.093%	0.810	**2990.974**
NJ	2431	-39980.181	2447	-39889.267	69.658%	0.012	3661.246
BIONJ	2431	-39963.162	2446	-39888.752	**69.659%**	**0.005**	3707.444

RDPII_218

	Best parsimony tree		Best likelihood tree		Hypervolume metrics		Execution
Methods	Pars.	Like.	Pars.	Like.	Mean	Std. Dev.	Times (s)
UPGMA	41716	-136739.845	42307	-134953.526	65.779%	0.602	5774.361
WPGMA	41750	-137306.291	42447	-135066.775	65.714%	0.582	5455.675
Complete	41643	-138581.502	42605	-135193.930	65.593%	0.686	4761.591
NJ	41488	-136297.204	42813	-134167.447	74.084%	0.245	3888.291
BIONJ	41488	-136267.570	42831	-134173.576	**74.086%**	**0.224**	**3862.010**

ZILLA_500

	Best parsimony tree		Best likelihood tree		Hypervolume metrics		Execution
Methods	Pars.	Like.	Pars.	Like.	Mean	Std. Dev.	Times (s)
UPGMA	16301	-82391.553	16354	-82001.417	55.409%	0.797	4800.333
WPGMA	16304	-81786.190	16314	-81776.387	57.339%	1.154	4796.018
Complete	16257	-87588.105	16293	-81632.727	59.588%	1.029	**4774.700**
NJ	16218	-81236.179	16310	-80966.641	69.064%	0.056	5974.231
BIONJ	16218	-81274.703	16310	-80966.525	**69.070%**	**0.055**	5998.784

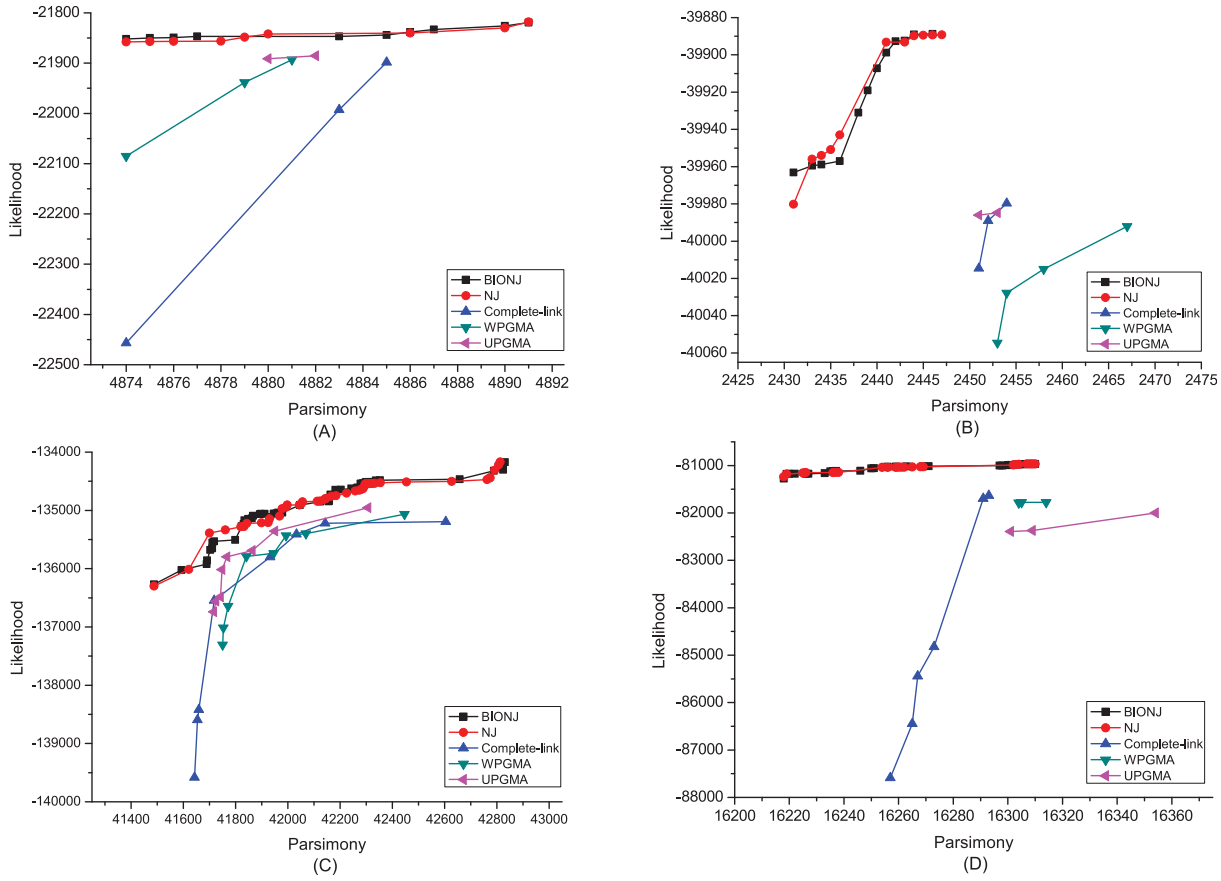


Figure 2: Pareto fronts generated for (A) *rbcL*_55, (B) *mtDNA*_186, (C) *RDPII*_218, and (D) *ZILLA*_500

According to the hypervolume values referenced in Table 3, there are significant differences between neighbor-joining methods and clustering techniques. NJ and BIONJ outperform the remaining approaches in terms of multiobjective performance and quality of the inferred trees in all the data sets. By analyzing results on *ZILLA*_500, we can observe that this improvement is more remarkable when complex data sets with high number of species are considered. Figure 3 displays graphically these growing differences in hypervolume values according with the complexity of the input dataset. While the selection of the tree-building method implies variations around 3.419% in hypervolume when analyzing the *rbcL*_55 dataset, this choice has a strong impact on *ZILLA*_500, giving rise to an average difference of 11.625% in hypervolumes for this dataset. These results reveal the importance of choosing a proper tree-building method based on realistic assumptions to address phylogenetic searches.

Therefore, we can state that neighbor-joining methods fit the considered biological data properly, allowing the algorithm to generate reliable phylogenies from the processed distance matrices for each data set. From a multiobjective point of view, hypervolume values show that BIONJ slightly improves NJ, and the comparison of standard deviations suggests that BIONJ can reduce the statistical variation from the average hypervolume values.

Concerning processing times, we can see that clustering methods require less times than neighbor-joining approaches in almost all data sets. This can be explained by analyzing the algorithmic complexities of each proposal. While modern implementations of clustering procedures require n^2 operations, neighbor-joining methods have an asymptotic complexity of $O(n^3)$ [9]. Among clustering methods, complete-linkage represents the most efficient approach. For the *RDPII*_218 dataset, processing times show that NJ and BIONJ outperform significantly clustering methods, due to

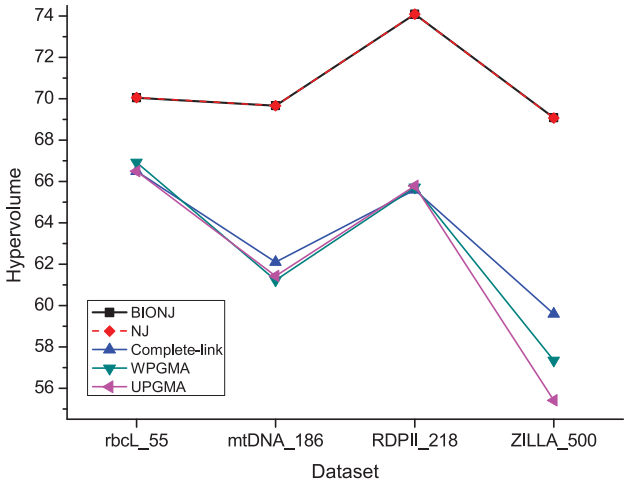


Figure 3: Hypervolume comparison

the fact that MO-FA can generate a significant number of non-dominated solutions, as seen in Figure 2. As the algorithm evolves, the amount of solutions which join the Pareto set grows, and the number of tasks which need to be performed (including calls to tree-building and evaluation procedures) is reduced, compensating the algorithmic complexity. This fact motivates improved times in comparison with methods based on traditional clustering analysis.

In conclusion, although clustering techniques require reduced complexities, BIONJ and NJ outperform UPGMA, WPGMA and complete-linkage in terms of multiobjective performance. Among clustering methods, complete-linkage shows the best processing times and is able to improve hypervolume values for $mtDNA\_186$ and $ZILLA\_500$. Concerning neighbor-joining methods, BIONJ attains the best hypervolume values in almost all data sets, and is able to perform efficient phylogenetic analyses for $RDPII\_218$. The model of variances and covariances included in BIONJ defines a statistically-consistent approach to generate phylogenetic trees, without introducing dramatic processing times with regard to NJ. Therefore, our experimental results suggest that a bioinspired multiobjective approach supported by a statistically reliable tree-building method can be useful to perform phylogenetic analyses on real data sets.

6.1 Comparisons with Other Proposals

In this subsection, we compare our swarm intelligence proposal with other author's approaches to phylogenetic reconstruction. By comparing with other multiobjective proposals and state-of-the-art single-criterion methods, we aim to evaluate the relevance of the proposal from a biological point of view. For this purpose, we consider the results achieved when adopting BIONJ as tree-building method.

In first place, we compare our approach with PhyloMOEA, a multiobjective evolutionary algorithm proposed by Cancino and Delbem to conduct phylogenetic analyses according to maximum parsimony and maximum likelihood [1]. Table 4 presents a comparison between the parsimony and likelihood scores achieved by our proposal with regard to the results reported by Cancino and Delbem under the $HKY85+\Gamma$ evolutionary model in [1]. According to this table, an algorithmic design which considers the knowledge gathered by all the individuals in the swarm gives as a result relevant solutions both from parsimony and likelihood perspectives, in comparison with a traditional evolutionary approach.

Table 4: Comparisons with PhyloMOEA

	MO-FA			
	Maximum parsimony tree		Maximum likelihood tree	
Dataset	Parsimony	Likelihood	Parsimony	Likelihood
rbcL_55	**4874**	-21852.01	4891	**-21819.19**
mtDNA_186	**2431**	-39963.16	2446	**-39888.75**
RDPII_218	**41488**	-136267.57	42831	**-134173.58**
ZILLA_500	**16218**	-81274.70	16310	**-80966.53**

	PhyloMOEA	
Dataset	Maximum parsimony score	Maximum likelihood score
rbcL_55	**4874**	-21889.84
mtDNA_186	2437	-39896.44
RDPII_218	41534	-134696.53
ZILLA_500	16219	-81018.06

Secondly, we assess our results through a comparison with two state-of-the-art biological methods: TNT [14], for maximum parsimony, and RAxML [28], for maximum likelihood reconstruction. In Table 5, we can observe that our swarm intelligence proposal generates high-quality phylogenies from both perspectives. With regard to parsimony scores, the comparison with TNT (column 2) shows that our bioinspired design can reach the reference scores provided by one of the most reliable tools for maximum parsimony. Additionally, in order to make possible a comparison with RAxML, we have performed new experiments by using the $GTR+\Gamma$ model to compute likelihood values. Columns 3 and 4 in Table 5 show that MO-FA can improve the parsimony and likelihood scores provided by RAxML in several data sets. Therefore, the proposed multiobjective approach represents a significant contribution from a biological perspective, obtaining quality results by combining multiobjective optimization techniques with collective intelligence.

Table 5: Comparisons with TNT and RAxML

	MO-FA		
	Maximum parsimony score	Maximum likelihood tree	
Dataset	Parsimony	Parsimony	Likelihood
rbcL_55	**4874**	**4890**	**-21783.48**
mtDNA_186	**2431**	**2448**	**-39868.74**
RDPII_218	**41488**	**42833**	-134082.27
ZILLA_500	**16218**	**16305**	**-80608.46**

	TNT	RAxML	
Dataset	Parsimony	Parsimony	Likelihood
rbcL_55	**4874**	4893	-21791.98
mtDNA_186	**2431**	2453	-39869.63
RDPII_218	**41488**	42894	**-134079.42**
ZILLA_500	**16218**	**16305**	-80623.50

7. CONCLUSIONS

In this paper we have reported a comparative study on the performance of several proposals for building phylogenies from distance matrices: UPGMA, WPGMA, complete-linkage, NJ, and BIONJ. These distance methods were applied to a multiobjective algorithm inspired by the collective behaviour of fireflies for inferring phylogenies, according to maximum parsimony and maximum likelihood. Experiments have been carried out on four real nucleotide data sets, and experimental results have been evaluated in terms of multiobjective performance and execution times.

Although processing times make clear the reduced complexities of clustering methods, neighbor-joining approaches can generate improved biological results, specially on complex data sets. The hypervolume metrics suggests that a configuration based on BIONJ leads the algorithm to the best overall behaviour, outperforming clustering methods even in processing times for the $RDPII\_218$ dataset. Additionally, comparisons with other multiobjective proposals and biological single-criterion methods show the relevance of our bioinspired metaheuristic.

As future research work, we will study the design of hybrid schemes based on MPI and OpenMP to parallelize MO-FA, with the aim of taking advantage of the characteristics of modern multicore clusters. In addition to this, we will undertake the development of other multiobjective metaheuristics and new parallel designs for inferring phylogenetic trees on data sets with a growing number of sequences.

8. ACKNOWLEDGMENTS

This work was partially funded by the Spanish Ministry of Economy and Competitiveness and the ERDF (European Regional Development Fund), under the contract TIN2012-30685 (BIO project). Sergio Santander-Jiménez is supported by the grant FPU12/04101 from the Spanish Government.

9. REFERENCES

[1] Cancino, W., and Delbem, A. C. B. A multi-criterion evolutionary approach applied to phylogenetic reconstruction. In *New Achievements in Evolutionary Computation*, P. Korosec, Ed. InTech, 2010, pp. 135–156.

[2] Cavalli-Sforza, L. L., and Edwards, A. W. F. Phylogenetic analysis: Models and estimation procedures. *American Journal of Human genetics 19*, 3 Pt 1 (1967), 223–257.

[3] Coelho, G. P., da Silva, A. E. A., and Zuben, F. J. V. Evolving phylogenetic trees: A multiobjective approach. In *BSB 2007* (2007), vol. 4643 of *LNCS (LNBI)*, Springer Verlag, pp. 113–125.

[4] Coello, C., Veldhuizen, D. V., and Lamont, G. Evolutionary algorithms for solving multi-objective problems. In *Genetic Algorithms and Evolutionary Computation*, vol. 5. Kluwer Academic Publishers, 2002.

[5] Congdon, C. B. An evolutionary algorithms approach for the study of natural evolution. In *Genetic and Evolutionary Computation Conference, GECCO 2002* (2002), pp. 1057–1064.

[6] Cotta, C., and Moscato, P. Inferring phylogenetic trees using evolutionary algorithms. In *Parallel Problem Solving From Nature VII* (2002), vol. 2439 of *LNCS*, Springer Verlag, pp. 720–729.

[7] Cotta, C., and Moscato, P. A memetic-aided approach to hierarchical clustering from distance matrices: application to gene expression clustering and phylogeny. *Biosystems 72*, 1-2 (2003), 75–97.

[8] Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution 17*, 6 (1981), 368–376.

[9] Felsenstein, J. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts, 2004.

[10] Fitch, W. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology 20*, 4 (1972), 406–416.

[11] Fogel, G. B. Evolutionary computation for the inference of natural evolutionary histories. *IEEE Connections 3*, 1 (2005), 11–14.

[12] Gascuel, O. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution 14*, 7 (1997), 685–695.

[13] Goëffon, A., Richer, J. M., and Hao, J. K. Progressive tree neighborhood applied to the maximum parsimony problem. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 5*, 1 (2008), 136–145.

[14] Goloboff, P. A., Farris, J. S., and Nixon, K. C. TNT, a free program for phylogenetic analysis. *Cladistics 24*, 5 (2008), 774–786.

[15] Gottlieb, J., Julstrom, B. A., Raidl, G. R., and Rothlauf, F. Prüfer numbers: A poor representation of spanning trees for evolutionary search. In *Genetic and Evolutionary Computation Conference, GECCO 2001* (2001), pp. 343–350.

[16] King, B. Step-wise clustering procedures. *J. Am. Statist. Assoc. 62*, 317 (1967), 86–101.

[17] Lemey, P., Salemi, M., and Vandamme, A.-M. *The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press, Cambridge, 2009.

[18] Lemmon, A. R., and Milinkovitch, M. C. The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation. *Proceedings of the National Academy of Sciences USA 99*, 16 (2002), 10516–10521.

[19] Lewis, P. O. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Molecular Biology and Evolution 15*, 3 (1998), 277–283.

[20] Matsuda, H. Construction of phylogenetic trees from amino acid sequences using a genetic algorithm. In *Proceedings of Genome Informatics Workshop* (1995), Universal Academy Press, pp. 19–28.

[21] Poladian, L. A GA for maximum likelihood phylogenetic inference using neighbour-joining as a genotype to phenotype mapping. In *Genetic and Evolutionary Computation Conference, GECCO 2005* (2005), pp. 415–422.

[22] Poladian, L., and Jermiin, L. Multi-objective evolutionary algorithms and phylogenetic inference with multiple data sets. *Soft Computing 10*, 4 (2006), 359–368.

[23] Saitou, N., and Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution 4*, 4 (1987), 406–425.

[24] Santander-Jimenez, S., and Vega-Rodriguez, M. A. A multiobjective proposal based on the firefly algorithm for inferring phylogenies. In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, EvoBIO 2013* (2013), vol. 7833 of *LNCS*, Springer Verlag, pp. 139–150.

[25] Skourikhine, A. Phylogenetic tree reconstruction using self-adaptive genetic algorithm. In *Proceedings of the 1st IEEE International Symposium on Bioinformatics and Biomedical Engineering* (2000), IEEE Computer Society, pp. 129–134.

[26] Sneath, P. H. A., and Sokal, R. R. *Numerical Taxonomy: the principles and practice of numerical classification*. W. H. Freeman, San Francisco, 1973.

[27] Sokal, R. R., and Michener, C. D. A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull. 38*, 2 (1958), 1409–1438.

[28] Stamatakis, A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics 22*, 21 (2006), 2688–2690.

[29] While, L., Hingston, P., Barone, L., and Huband, S. A faster algorithm for calculating hypervolume. *IEEE Transactions on Evolutionary Computation 10*, 1 (2006), 29–38.

[30] Wiens, J. J., and Servedio, M. R. Phylogenetic analysis and intraspecific variation: performance of parsimony, likelihood, and distance methods. *Systematic Biology 47*, 2 (1998), 228–253.

[31] Yang, X. S. Firefly algorithm, stochastic test functions and design optimisation. *Int. J. Bio-Inspired Computation 2*, 2 (2010), 78–84.