A Multiobjective Proposal Based on the Firefly Algorithm for Inferring Phylogenies

Sergio Santander-Jiménez and Miguel A. Vega-Rodríguez

University of Extremadura,
Department of Technologies of Computers and Communications,
ARCO Research Group.
Escuela Politécnica. Campus Universitario s/n, 10003. Cáceres, Spain
{sesaji,mavega}@unex.es

Abstract. Recently, swarm intelligence algorithms have been applied successfully to a wide variety of optimization problems in Computational Biology. Phylogenetic inference represents one of the key research topics in this area. Throughout the years, controversy among biologists has arisen when dealing with this well-known problem, as different optimality criteria can give as a result discordant genealogical relationships. Current research efforts aim to apply multiobjective optimization techniques in order to infer phylogenies that represent a consensus between different principles. In this work, we apply a multiobjective swarm intelligence approach inspired by the behaviour of fireflies to tackle the phylogenetic inference problem according to two criteria: maximum parsimony and maximum likelihood. Experiments on four real nucleotide data sets show that this novel proposal can achieve promising results in comparison with other approaches from the state-of-the-art in Phylogenetics.

Keywords: Swarm Intelligence, Multiobjective Optimization, Phylogenetic Inference, Firefly Algorithm.

1 Introduction

Bioinformatics aims to address problems that imply the processing of a growing number of biological data by means of computational techniques. Most of these problems cannot be tackled by using exhaustive searches because of their NP-hard complexity. Additionally, these biological problems can be addressed from several, conflicting perspectives. Recent research works try to overcome such limitations by applying multiobjective metaheuristics [1]. Their main goal is to generate a set of Pareto solutions that represent a compromise between different criteria, by optimizing simultaneously two or more objective functions [2].

One of the key problems in Computational Biology is the reconstruction of ancestral genealogical relationships among species, phylogenetic inference [3]. Phylogenetic procedures take as input molecular characteristics of organisms, such as nucleotide sequences represented by using an alphabet $\Sigma = \{A, C, G, T\}$. By analyzing these sequences, we get a mathematical structure that describes a

L. Vanneschi, W.S. Bush, and M. Giacobini (Eds.): EvoBIO 2013, LNCS 7833, pp. 141–152, 2013.

hypothesis about the evolutionary history of these species, the phylogenetic tree. Input species represent the results of the evolutionary process and are located at the leaves of the tree. Hypothetical ancestors are represented by internal nodes, and ancestor-descendant relationships are modelled by branches.

In the literature we can find a variety of optimality criteria for inferring phylogenies, such as maximum parsimony, maximum likelihood and distance methods [3]. However, by using a specific criterion, the resulting phylogenies can be radically different to the trees generated by other criteria, inferring discordant genealogical relationships [4]. This fact motivates that biologists are forced to use several single-criterion software to analyze complex data sets, and publish results that make clear these conflicts. By means of multiobjective optimization, we try to support a complementary view of phylogenetic inference according to multiple criteria, with the aim of generating a set of phylogenetic trees that represent a consensus between different points of view.

In this paper we propose a multiobjective adaptation of the novel Firefly Algorithm (FA) for inferring phylogenetic trees attending to the parsimony and likelihood principles. We have chosen this swarm intelligence algorithm due to the promising results reported for a variety of problems, overcoming other bioinspired proposals [5]. In order to assess the performance of this Multiobjective Firefly Algorithm (MO-FA), we have carried out experiments on four nucleotide data sets, applying the hypervolume metrics [2], and comparing results with other authors' multiobjective proposals and popular single-criterion methods.

This paper is organized as follows. In the next section, we introduce a short review on other bioinspired proposals for inferring phylogenetic trees. In Section 3, we detail the basis of phylogenetic methods based on distances, parsimony and likelihood. Section 4 explains the details about MO-FA and discusses how to adapt it to multiobjective phylogenetic inference. Experimental results are presented and explained in Section 5, introducing comparisons with other proposals. Finally, Section 6 summarizes some conclusions and future research lines.

2 Related Work

Throughout the years, several bioinspired proposals have been published with the aim of carrying out phylogenetic analyses on data sets with a growing complexity. In such data sets, exhaustive searches cannot be applied due to the huge number of possible phylogenetic topologies, which increases in an exponential way with the number of species [3]. In this section, we summarize several bioinspired approaches to Phylogenetics proposed by other authors.

The first bioinspired proposals for inferring phylogenies were reported by Matsuda [6] and Lewis [7], in 1995 and 1998, respectively. Following this line, other researchers published new approaches for tackling the phylogenetic inference problem. We can highlight the work of Lemmon and Milinkovitch, who published a multipopulation genetic algorithm for maximum likelihood reconstruction [8], and Congdon, who developed an evolutionary algorithm for maximum parsimony analyses [9]. Recently, the basis of bioinspired computing can be found in some

methodologies included in popular biological methods [10], [11]. One of most important questions that arises when developing these strategies is how to represent individuals in the population. Cotta and Moscato studied several direct and indirect representations, observing different advantages and disadvantages [12]. On the other hand, Poladian proposed in [13] the use of distance matrices and the Neighbour-Joining method as a genotype-phenotype mapping, applied to maximum likelihood. His proposal achieved promising results with regard to other popular heuristic-based approaches.

Recent research trends suggest the use of multiobjective optimization techniques applied to Phylogenetics. This line was defined to overcome the difficulties that arise when using these previous approaches, as several sources of evidence and different optimality criteria can give as a result conflicting tree topologies. In 2006, Poladian and Jermiin developed the first multiobjective algorithm applied to phylogenetic reconstruction [14]. Afterwards, an immune-inspired multiobjective proposal for inferring phylogenies by the minimal evolution and mean-squared error criteria was proposed by Coelho et al. [15]. Finally, Cancino and Delbem published a multiobjective genetic algorithm for maximum parsimony and maximum likelihood reconstruction, PhyloMOEA [16].

Following this last line of research, in this paper we introduce a new multiobjective bioinspired approach for inferring phylogenies that represent a consensus between maximum parsimony and maximum likelihood.

3 Approaches for Inferring Phylogenies

In this section we introduce the basis of different phylogenetic methods whose characteristics will be considered in our proposal: distance methods, maximum parsimony and maximum likelihood approaches.

3.1 Distance-Based Methods

Distance-based methods [3] were proposed with the aim of inferring phylogenetic trees by processing some distance measures among species. Despite their simplicity, these methods are very popular due to the low amount of biological information lost when modelling the evolutionary process [3]. Furthermore, these approaches can lay the foundations for more complex phylogenetic searches. Distance methods generate a symmetric matrix M of $N \times N$ dimensions, where N is the number of species in input data. Given two species i and j, $M[i, j]$ contains the evolutionary distance between them. A distance measure can be computed in several ways, such as by considering the number of different characteristics found in molecular sequences, or by using statistical methods [13].

These approaches process M to generate a phylogenetic topology $T = (V, E)$, where V represents the nodes in the tree, and E contains branches modelling ancestral relationships, as well as evolutionary distances between related organisms, given by branch length values. Figure 1 shows an example of distance-based phylogenetic reconstruction. Among the variety of distance methods that can be

	Seal	Whale	Rat	Mouse	Human	Turkey	Frog	Minnow	Carp	Bull
Seal	0	0.20	0.30	0.31	0.49	0.45	0.40	0.50	0.49	0.19
Whale	0.20	0	0.24	0.25	0.43	0.39	0.34	0.44	0.43	0.19
Rat	0.30	0.24	0	0.15	0.45	0.41	0.36	0.46	0.45	0.29
Mouse	0.31	0.25	0.15	0	0.46	0.42	0.37	0.47	0.46	0.30
Human	0.49	0.43	0.45	0.46	0	0.34	0.43	0.53	0.52	0.48
Turkey	0.45	0.39	0.41	0.42	0.34	0	0.39	0.49	0.48	0.44
Frog	0.40	0.34	0.36	0.37	0.43	0.39	0	0.30	0.29	0.39
Minnow	0.50	0.44	0.46	0.47	0.53	0.49	0.30	0	0.21	0.49
Carp	0.49	0.43	0.45	0.46	0.52	0.48	0.29	0.21	0	0.48
Bull	0.19	0.19	0.29	0.30	0.48	0.44	0.39	0.49	0.48	0

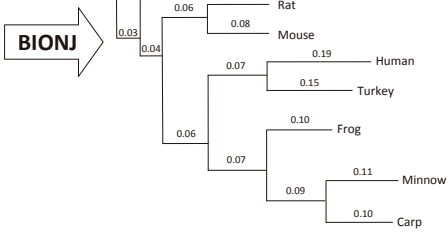

Fig. 1. An example of the BIONJ algorithm considering ten species

found in the literature, Neighbour-Joining (NJ) [17] is commonly used as an understandable way to introduce researchers into this methodology.

NJ verifies M iteratively, selecting the pair (s_1, s_2) of species that represent the closest neighbours according to the distance measures. From s_1 and s_2, the method infers their common ancestor c and includes (c, s_1, s_2) in a partial phylogeny. Entries related to s_1 and s_2 in M are replaced by c and distance values are updated. These steps are repeated until all entries in M have been processed and a complete phylogenetic topology has been inferred. In this work, we will use the BIONJ method, an extension to NJ developed by Gascuel. This proposal improves NJ performance by considering a model of variances and covariances estimated from evolutionary distances. A detailed explanation of this algorithm can be found in [18].

3.2 Maximum Parsimony Approaches

Ockham's razor has inspired a wide variety of approaches to resolve optimization problems. Maximum parsimony methods try to apply this well-known principle to Phylogenetics. Parsimony-based approaches are defined to infer those evolutionary histories that minimize the amount of molecular changes needed to explain the observed data. Given a phylogenetic tree $T = (V, E)$ that describes an evolutionary history from a set of N nucleotide sequences composed by K sites, we compute the parsimony value of T by using the following equation [19]:

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

where $(a, b) \in E$ represents an ancestor-descendant relationship between the nodes a and b, a_i and b_i the states corresponding to the ith site on molecular sequences for a and b, and $C(a_i, b_i)$ the cost of evolving from the state a_i to b_i.

Those trees that minimize Equation 1 will be preferred as they imply a simpler hypothesis to the evolution of the input species. In this work we will consider

Fitch's proposal to assign ancestral sequences to internal nodes and evaluate trees according to the maximum parsimony criterion [20].

3.3 Maximum Likelihood Approaches

Likelihood-based approaches to Phylogenetics were proposed to infer the most likely evolutionary history of the organisms under review by using complex evolutionary models. Evolutionary models provide substitution matrices that define mutation probabilities at nucleotide level. By considering such models, the topology of the phylogenetic tree and branch lengths values, these methods compute statistically consistent phylogenies, according to the likelihood measurement.

Let $T = (V, E)$ be a phylogenetic tree, m an evolutionary model, and D a set of K-site nucleotide sequences representing the observed data. We can calculate the likelihood of T as follows [3]:

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

where r_i is defined as the mutation probability for the ith site, t_j as evolutionary times given by branch $j \in E$, and n_{ij} as the number of state changes that can be found on site i between the nodes related by j.

These approaches search for those phylogenies that maximize the likelihood function, due to the fact that maximum likelihood topologies would represent the most likely evolutionary hypotheses. We will compute likelihood values by using the Felsenstein proposal [21] under the $HKY85 + \Gamma$ model [3].

4 Multiobjective Firefly Algorithm

Recently, a novel swarm intelligence algorithm inspired by the bioluminescence of fireflies was proposed by Yang [5]. The Firefly Algorithm uses concepts like brightness and attractiveness to resolve optimization problems by using collective intelligence. Fireflies behaviour is governed by a communication system based on flashing lights which allows them to attract other fireflies and to warn predators about their toxicity. Attractiveness depends on the light intensity, the distance between fireflies and the light absorption by environment. Brighter fireflies will attract less bright fireflies to their position. FA models this behaviour by considering that the light intensity of a firefly will depend on the quality of its related solution. Brighter fireflies will be associated to better solutions to the problem, so firefly population will move towards high-quality solutions.

In this study, we propose to introduce multiobjective optimization techniques to FA. For this purpose, we need to distinguish brighter fireflies to less bright fireflies in this new context. To resolve this issue, we apply *dominance*. Given two solutions x and y, we state that x dominates y if and only if x has better or equal scores than y in all objective functions and, at least, x is better in one of them. In this way, given two fireflies u and v with solutions X_u and X_v, respectively, u will be brighter than v if and only if X_u dominates X_v.

In order to adapt this Multiobjective Firefly Algorithm to phylogenetic inference, we will use the distance-based methodology proposed by Poladian in [13]. We will introduce distance matrices to model the attraction process and apply BIONJ to reconstruct the resulting phylogenetic trees. Algorithm 1 shows MO-FA pseudocode. This algorithm takes as input the following parameters:

1. swarmSize. Population size.
2. maxGenerations. Number of generations.
3. β_0. Attractiveness factor.
4. γ. Environment absorption coefficient.
5. α. Randomization factor.

Algorithm 1. MO-FA Pseudocode

```
1: X ← initializeAndEvaluatePopulation(swarmSize,dataset)
2: ParetoFront ← 0
3: i ← 0
4: while i < maxGenerations do
5:    for j = 1 to swarmSize do
6:       for k = 1 to swarmSize do
7:          /* If X[k] dominates X[j], X[j] will move towards X[k] */
8:          if X[k] ≻ X[j] then
9:             /* Compute distance from X[k] to X[j] and apply attraction formula */
10:            r_jk ← ||X[j] − X[k]|| = sqrt(Σ_{n=1}^{N} Σ_{m=1}^{n} (X[j].M[n,m] − X[k].M[n,m])^2)
11:            for each position m,n (n < m) in X[j] distance matrix do
12:               X[j].M[m,n] ← X[j].M[m,n] + β0 e^{−γ r_jk^2} (X[k].M[m,n] − X[j].M[m,n]) +
                  α(rand[0,1] − 1/2)
13:            end for
14:            X[j].T ← computeBIONJ(X[j].M)
15:            X[j] ← setParsimonyAndLikelihoodScores(X[j].T, dataset)
16:         end if
17:      end for
18:   end for
19:   X ← optimizeExtremeSolutions(X)
20:   ParetoFront ← saveSolutions(X, ParetoFront)
21:   i ← i + 1
22: end while
```

Initializing the Swarm. Each firefly in the swarm will be related to a distance matrix and the corresponding phylogenetic topology. We have used the matrix and tree templates provided by the C++ libraries for Bioinformatics, BIO++ [22]. Initial trees are selected from a repository of 1000 phylogenetic topologies generated by bootstrap techniques, 500 of them by using maximum parsimony, and the remaining 500 by maximum likelihood. In addition to this, BIO++ is used to configure the parameters of the evolutionary model.

From these topologies, initial scores and distance matrices are computed and assigned to individuals in the population. BIONJ will be used to generate phylogenetic topologies from the updated matrices during the course of the algorithm.

MO-FA Main Loop. After the firefly population has been initialized, MO-FA main loop takes place (lines 4-22 in Algorithm 1). Each dominated firefly will be modified according to the brightness and attractiveness system. For this purpose, entries in the distance matrix will be updated in order to move dominated fireflies towards the brightest ones. Firstly, given two fireflies u, v with solutions X_u and X_v, if X_u is dominated by X_v, we compute the distance r_{uv} between u and v as follows (line 10):

$$r_{uv} \leftarrow \|X_u - X_v\| = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{i}(X_u.M[i,j] - X_v.M[i,j])^2} \tag{3}$$

where $M[i,j]$ denotes the distance between species i and j. According to the analyzed dataset, distance values between fireflies can be significantly different, so we normalize resulting distances to a specific range, [0,10].

In second place, we use r_{uv} to compute the new distance matrix, updating each entry $M[i,j]$ according to MO-FA movement formula. Given the attractiveness β_0, the environment absorption coefficient γ and a randomization factor α, the updated distance between two species i and j is given by (line 12):

$$\begin{aligned} X_u.M[i,j] =& X_u.M[i,j] + \beta_0 e^{-\gamma r_{uv}^2}(X_v.M[i,j] - X_u.M[i,j]) \\ &+ \alpha(rand[0,1] - \frac{1}{2}) \end{aligned} \tag{4}$$

The second term in Equation 4 denotes how $X_u.M[i,j]$ will move towards $X_v.M[i,j]$, taking into account β_0 and γ. The third term introduces a randomization factor to the movement of fireflies which helps to maintain the population diversity, where *rand* represents a random number generator. By combining the knowledge provided by the swarm with randomness, we can address the search for quality phylogenetic topologies in undiscovered regions of the tree space.

Once the distance matrix has been updated, the BIONJ algorithm generates the new phylogenetic tree according to the new distances computed by using Equation 4. Resulting topologies will be evaluated then according to the maximum parsimony and maximum likelihood criteria (line 15).

Final Steps. After all fireflies have been processed, we apply an optimization step in order to introduce additional knowledge provided by well-known heuristic-based searches. For this reason, extreme points in Pareto front will be optimized by applying a local search procedure based on the Parametric Progressive Tree Neighbourhood (PPN) proposed by Goëffon et al. [19]. PPN neighbours will be evaluated attending to the dominance concept, with the aim of improving maximum parsimony and maximum likelihood scores. Additionally, a gradient method is applied to improve branch length values. Fireflies will learn from these new solutions, allowing the swarm to improve the quality of Pareto solutions.

At the end of the current generation, the Pareto nondominated set is updated with the current best phylogenetic trees, and a new generation takes place.

After *maxGenerations*, the Pareto set will be composed by those phylogenetic trees that suppose a compromise between the parsimony and likelihood principles.

5 Experimental Methodology and Results

In this section we explain the experimental methodology we have followed to assess the performance of the proposal, evaluating and comparing our biological results with different approaches for inferring phylogenies. We have used a well-known quality indicator in multiobjective optimization, the hypervolume metrics [2], to evaluate the quality of the inferred Pareto solutions. Hypervolume defines the size of the search space covered by our solutions, bounded by two reference points (ideal and nadir). Metaheuristics that maximize hypervolume will be preferred over other proposals in a multiobjective context. In Table 1, we show the reference points we have used to compute hypervolume values.

Table 1. Hypervolume metrics. Reference points

	Ideal Point		Nadir 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

In order to configure our proposal, we have considered a variety of values for the three main parameters of the algorithm, β_0, γ and α. The remaining parameters, *maxGenerations* and *swarmSize*, have been configured taking into account additional experiments and other authors' proposals [16]. The different values we have studied for β_0, γ and α can be found in Table 2. We have chosen by experimentation the configuration that allows MO-FA to maximize hypervolume values. Table 3 shows the resulting values for input parameters.

Table 2. Configuring parameters

Parameter	Values
β_0	{0.05, 0.1, 0.25, 0.5, 0.75, 1}
γ	{0.05, 0.1, 0.25, 0.5, 0.75, 1}
α	{0.05, 0.1, 0.25, 0.5, 0.75, 0.9}

Table 3. MO-FA input parameters

Parameter	Final value
maxGenerations	100
swarmSize	100
β_0	1
γ	0.5
α	0.05

Experiments have been carried out on four nucleotide data sets [16] using the $HKY85+\Gamma$ model: *rbcL*_55, 55 sequences of 1314 nucleotides per sequence of the rbcL gene from green plants, *mtDNA*_186, 186 sequences of 16608 nucleotides per sequence from human mitochondrial DNA, *RDPII*_218, 218 sequences of

Table 4. Experimental results

	Pareto trees	Maximum parsimony tree		Maximum likelihood tree		Best hypervolume tree		Hypervolume metrics	
Dataset		Pars.	Like.	Pars.	Like.	Pars.	Like.	Mean	Std. Dev.
rbcL_55	8	4874	-21849.36	4892	-21819.04	4882	-21830.76	70.06%	0.06428
mtDNA_186	12	2431	-39961.98	2448	-39888.58	2439	-39903.13	69.67%	0.01251
RDPII_218	39	41488	-136340.73	42833	-134169.03	41745	-135409.63	74.01%	0.33689
ZILLA_500	28	16218	-81613.47	16309	-80966.58	16221	-81212.39	69.06%	0.05880

4182 nucleotides per sequence from prokaryotic RNA, and *ZILLA_500*, 500 sequences of 759 nucleotides per sequence from rbcL plastid gene.

For each dataset, we have performed 30 independent analyses to assess the statistical relevance of the proposal. In Table 4, we summarize the results corresponding to the execution which achieved the closest score to the mean hypervolume value. Columns 3-4 and 5-6 show parsimony and likelihood values for the extreme points in Pareto front. Additionally, parsimony and likelihood scores for the non-extreme solution that contributed most to the overall hypervolume are given by Columns 7-8. Finally, mean hypervolume values and standard deviations are indicated in Columns 9-10. According to this table, the hypervolume metrics suggest that MO-FA gets significant Pareto solutions for all data sets, covering over 69% of the space bounded by reference points. Pareto fronts for each dataset can be found in Figure 2.

5.1 Comparisons with Other Proposals

In order to assess the quality of the inferred phylogenetic trees, in this subsection we compare MO-FA with other authors' multiobjective metaheuristics and popular biological methods for inferring phylogenies.

In first place, we introduce in Table 5 a comparison with PhyloMOEA, a multiobjective algorithm for maximum parsimony and maximum likelihood phylogenetic reconstruction. In this table, we show parsimony and likelihood scores for our maximum parsimony and maximum likelihood trees and compare them with the best values reported by Cancino and Delbem's proposal in [16], using $HKY85 + \Gamma$. Results suggest a significant improvement with regard to PhyloMOEA in all data sets, inferring phylogenetic trees that overcome the best scores provided by other authors' multiobjective approaches. As swarm intelligence allows the inference process to take into account knowledge provided by different fireflies, a better exploration of the tree space can be performed, dominating the results achieved by classical multiobjective evolutionary algorithms.

Secondly, we compare MO-FA with two well-known single-criterion methods from the state-of-the-art: TNT [10], for maximum parsimony reconstruction, and RAxML [11], for maximum likelihood. In Table 6 we can find the best parsimony scores achieved by MO-FA and TNT, as well as parsimony and likelihood scores for the maximum likelihood trees inferred by MO-FA and RAxML. Attending to parsimony, our proposal achieves the reference scores provided by TNT. With regard to likelihood comparison, as new versions of RAxML does not include

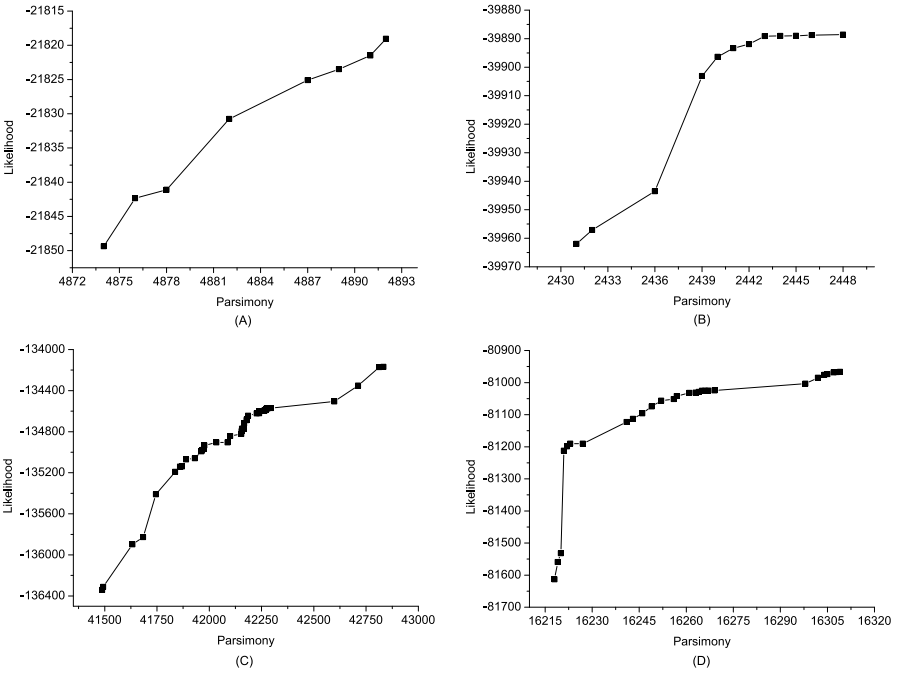

Fig. 2. Pareto fronts for *rbcL_55*(A), *mtDNA_186*(B), *RDPII_218*(C) and *ZILLA_500*(D)

Table 5. Comparing MO-FA with PhyloMOEA

	MO-FA			
	Best parsimony tree		Best likelihood tree	
Dataset	Parsimony	Likelihood	Parsimony	Likelihood
rbcL_55	**4874**	-21849.36	4892	**-21819.04**
mtDNA_186	**2431**	-39961.98	2448	**-39888.58**
RDPII_218	**41488**	-136340.73	42833	**-134169.03**
ZILLA_500	**16218**	-81613.47	16309	**-80966.58**
	PhyloMOEA			
Dataset	Best parsimony score		Best likelihood score	
rbcL_55	**4874**		-21889.84	
mtDNA_186	2437		-39896.44	
RDPII_218	41534		-134696.53	
ZILLA_500	16219		-81018.06	

Table 6. Comparing MO-FA with TNT and RAxML

	MO-FA		
	Best parsimony score	Best likelihood tree	
Dataset	Parsimony	Parsimony	Likelihood
rbcL_55	**4874**	**4890**	**-21789.27**
mtDNA_186	**2431**	**2451**	**-39869.29**
RDPII_218	**41488**	**42813**	-134089.91
ZILLA_500	**16218**	**16305**	**-80610.86**
	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

$HKY85 + \Gamma$, we have carried out new experiments using the $GTR + \Gamma$ evolutionary model. Under this model, our likelihood topologies dominate RAxML's trees for *rbcL_55*, *mtDNA_186* and *ZILLA_500*, and improve significantly the parsimony value for *RDPII_218*. Therefore, we can suggest that a multiobjective swarm intelligence scheme allows us to obtain a meaningful performance in comparison with two of the most powerful tools for phylogenetic inference.

6 Conclusions and Future Work

We have introduced in this paper a multiobjective approach based on the collective behaviour of fireflies for tackling the phylogenetic inference problem according to two well-known criteria: maximum parsimony and maximum likelihood. In order to model fireflies' behaviour, we have used a distance-based methodology supported by the BIONJ algorithm, where distance matrices are computed and processed to generate new phylogenetic topologies. Experiments on four public nucleotide data sets show that this swarm intelligence proposal can achieve significant performance in comparison with other multiobjective evolutionary algorithms and state-of-the-art biological methods, inferring a set of trade-off phylogenetic trees by considering the parsimony and likelihood principles.

As future work, we will introduce this distance-based methodology and individual representation into a previous swarm intelligence algorithm for inferring phylogenies, Multiobjective Artificial Bee Colony (MOABC) [23], with the aim of making possible a fair comparison between MOABC and MO-FA. The reason why we need to study such step is because performing this comparison without taking into account the same experimental conditions can give as a result biased conclusions. Additionally, other distance methods besides BIONJ will be studied, in order to assess which one can lead MO-FA to improved performances. Finally, we will apply parallel computing to improve the efficiency of the proposal by exploiting modern hardware architectures.

Acknowledgment. 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). Thanks to the Fundación Valhondo Calaff for the financial support offered to Sergio Santander-Jiménez.

References

1. Handl, J., Kell, D., Knowles, J.: Multiobjective Optimization in Computational Biology and Bioinformatics. IEEE Transactions on Computational Biology and Bioinformatics 4(2), 289–292 (2006)
2. Zitzler, E., Thiele, L.: Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach. IEEE Transactions on Evolutionary Computation 3(4), 257–271 (1999)
3. Felsenstein, J.: Inferring phylogenies. Sinauer Associates, Sunderland (2004) ISBN: 0-87893-177-5
4. Wiens, J.J., Servedio, M.R.: Phylogenetic analysis and intraspecific variation: performance of parsimony, likelihood, and distance methods. Systematic Biology 47(2), 228–253 (1998)
5. Yang, X.-S.: Firefly Algorithm, Stochastic Test Functions and Design Optimisation. Int. J. Bio-Inspired Computation 2(2), 78–84 (2010)
6. Matsuda, H.: Construction of phylogenetic trees from amino acid sequences using a genetic algorithm. In: Proceedings of Genome Informatics Workshop, pp. 19–28. Universal Academy Press (1995)

7. Lewis, P.O.: A Genetic Algorithm for Maximum-Likelihood Phylogeny Inference Using Nucleotide Sequence Data. Mol. Biol. Evol. 15(3), 277–283 (1998)
8. Lemmon, A.R., 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), 10516–10521 (2002)
9. Congdon, C.: GAPHYL: An evolutionary algorithms approach for the study of natural evolution. In: Genetic and Evolutionary Computation Conference, GECCO 2002, pp. 1057–1064 (2002)
10. Goloboff, P.A., Farris, J.S., Nixon, K.C.: TNT, a free program for phylogenetic analysis. Cladistics 24, 774–786 (2008)
11. Stamatakis, A.: RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. Bioinformatics 22(21), 2688–2690 (2006)
12. Cotta, C., Moscato, P.: Inferring Phylogenetic Trees Using Evolutionary Algorithms. In: Guervós, J.J.M., Adamidis, P., Beyer, H.-G., Fernández-Villacañas, J.-L., Schwefel, H.-P. (eds.) PPSN VII. LNCS, vol. 2439, pp. 720–729. Springer, Heidelberg (2002)
13. 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, pp. 415–422 (2005)
14. Poladian, L., Jermiin, L.: Multi-Objective Evolutionary Algorithms and Phylogenetic Inference with Multiple Data Sets. Soft Computing 10(4), 359–368 (2006)
15. Coelho, G.P., da Silva, A.E.A., Von Zuben, F.J.: Evolving Phylogenetic Trees: A Multiobjective Approach. In: Sagot, M.-F., Walter, M.E.M.T. (eds.) BSB 2007. LNCS (LNBI), vol. 4643, pp. 113–125. Springer, Heidelberg (2007)
16. Cancino, W., Delbem, A.C.B.: A Multi-Criterion Evolutionary Approach Applied to Phylogenetic Reconstruction. In: Korosec, P. (ed.) New Achievements in Evolutionary Computation, pp. 135–156. InTech (2010) ISBN: 978-953-307-053-7
17. Saitou, N., Nei, M.: The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. Molecular Biology and Evolution 4(4), 406–425 (1987)
18. Gascuel, O.: BIONJ: An Improved Version of the NJ Algorithm Based on a Simple Model of Sequence Data. Molecular Biology and Evolution 14(7), 685–695 (1997)
19. Goëffon, A., Richer, J.M., Hao, J.K.: Progressive Tree Neighborhood Applied to the Maximum Parsimony Problem. IEEE/ACM Transactions on Computational Biology and Bioinformatics 5, 136–145 (2008)
20. Fitch, W.: Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. Systematic Zoology 20(4), 406–416 (1972)
21. Felsenstein, J.: Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. Journal of Molecular Evolution 17, 368–376 (1981)
22. Dutheil, J., Gaillard, S., Bazin, E., Glémin, S., Ranwez, V., Galtier, N., Belkhir, K.: Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. BMC Bioinformatics 7, 188–193 (2006)
23. Santander-Jiménez, S., Vega-Rodríguez, M.A., Gómez-Pulido, J.A., Sánchez-Pérez, J.M.: Inferring Phylogenetic Trees Using a Multiobjective Artificial Bee Colony Algorithm. In: Giacobini, M., Vanneschi, L., Bush, W.S. (eds.) EvoBIO 2012. LNCS, vol. 7246, pp. 144–155. Springer, Heidelberg (2012)