A Memetic Algorithm Based on an NSGA-II Scheme for Phylogenetic Tree Inference

Manuel Villalobos-Cid, Márcio Dorn, Rodrigo Ligabue-Braun, and Mario Inostroza-Ponta

***Abstract*—Phylogenetic inference allows building a hypothesis about the evolutionary relationships between a group of species, which is usually represented as a tree. The phylogenetic inference problem can be seen as an optimization problem, searching for the most qualified tree among all the possible topologies according to a selected criterion. These criteria can be based on different principles. Due to the combinatorial number of possible topologies, diverse heuristics and meta-heuristics have been proposed to find approximated solutions according to one criterion. However, these methods may result in several phylogeny trees which could be in conflict with one another. In order to deal with this problem, models based on multiobjective optimization with different configurations have been used. In this paper, we propose an *ad-hoc* multiobjective memetic algorithm (MO-MA) to infer phylogeny using two objectives: 1) maximum parsimony and 2) likelihood. Several population operators and local search strategies are proposed and evaluated in order to measure their contribution to the algorithm. Additionally, we perform a comparison among different configurations and tree rearrangement strategies. The results show that the proposed MO-MA is able to identify a Pareto set of solutions that include new trees which were nondominated by solutions from the current state of the art single-objective optimization tools. Furthermore, the MO-MA improves the results presented in the literature for multiobjective approaches in all of the studied data sets. These results make our proposal a good alternative for phylogenetic inference.**

***Index Terms*—Memetic algorithm, multiobjective optimization, phylogenetic inference.**

Manuscript received November 15, 2016; revised July 27, 2018; accepted November 15, 2018. Date of publication November 28, 2018; date of current version October 1, 2019. This work was supported by Microsoft through the Microsoft Azure for Research Award. The work of M. Villalobos-Cid and M. Inostroza-Ponta was supported in part by CeBIB under Grant FB00001, in part by CITIAPS under Grant PMI USA1204, and in part by DICYT-VRIDEI, USACH under Grant 061619IP. The work of M. Dorn and R. Ligabue-Braun was supported in part by FAPERGS (PRONUPEQ) under Grant 002021-25.51/13 and Grant 16/2551-0000520-6, and in part by MCT/CNPq under Grant 311022/2015-4. *(Corresponding author: Mario Inostroza-Ponta.)*

M. Villalobos-Cid and M. Inostroza-Ponta are with the Departamento de Ingeniería Informática, Facultad de Ingeniería, Centre for Biotechnology and Bioengineering, Universidad de Santiago de Chile, Santiago 8320000, Chile (e-mail: mario.inostroza@usach.cl).

M. Dorn is with the Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre 15064, Brazil.

R. Ligabue-Braun is with the Department of Pharmaceutical Sciences, Federal University of Health Sciences of Porto Alegre, Porto Alegre 90050-170, Brazil.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org provided by the author. This consists of a PDF file containing further details of the algorithm implementation and results obtained. This material is 537 KB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TEVC.2018.2883888

I. Introduction

ONE OF the major challenges of bioinformatics is the design of efficient and robust algorithms to deal with problems that arise in this field [1]. In particular, the phylogenetic inference problem allows the inference of a hypothesis about the evolutionary history between a group of organisms (usually called taxa or tips), using structural and functional information of molecules and how they change over time [2]. Phylogenetic analysis has contributed in several scientific fields such as bio-geography, medicine, chemistry, forensics, systematic biology, epidemiology, and paleontology [3]. For example, the publication of the tree of life proposed by Hinchliff *et al.* [4] shows the evolutionary relationships among 2.3 million organisms. This paper also shows the large volume of data that is available for phylogenetic inference, including multiple data sets, biological sources, and different species.

One way to represent the evolutionary relationships is using phylogenetic trees, so the problem of phylogenetic inference can be treated as an optimization problem. Then, it is required to find the most qualified inferred tree among all the possible topologies according to a selected criterion, which can be based on different principles such as minimum evolution, least squares, maximum parsimony, and likelihood. All of them can result in different phylogenetic trees. The search for the optimal tree using approaches based on exhaustive algorithms is infeasible, due to the combinatorial number of possible topologies. This problem has been classified in computational theory as an NP-hard problem [5].

The first approaches for this problem considered only one optimization criterion and they were based on evolutionary and bio-inspired algorithms [6], [7]. They allowed to find approximate solutions to problem instances with large-scale volume data (see [7] and references therein). Nevertheless, some problems remained unsolved such as the bias associated with the optimal criterion applied [8], type of data, and evolutionary model selected [9]. Probabilistic algorithms provided a solution to these difficulties generating a consensus (CS) tree. However, they require prior knowledge of the weight of each biological source and objectives without conflict with each other. The main shortcoming of this methodology would be to discard solutions with biological meaning [5].

Handl *et al.* [10] highlighted the advantages of applying multiobjective optimization in bioinformatics and computational biology problems. They identified its benefits compared to the use of single objective methods: minimization of the probability of stagnation in local minima and areas without

gradients, reduction of noise effect from the data, and incorporation of multiple sources which are in conflict with each other.

The most recent methods for phylogenetic inference using trees are based on multiobjective optimization [1], [5], [7], [9], [11]–[17]. These population-based evolutionary strategies consider a wide range of meta-heuristics, different methods to build initial populations, and diverse genetic operators for crossover and mutation. However, only a few of them [14], [15], [18] describe details about the performance of rearrangement strategies over genetic operations and how the algorithms are parameterized.

In this paper, we propose a memetic algorithm based on a nondominated sorting genetic algorithm II (NSGA-II) to infer phylogeny [19] using two objectives: 1) maximum parsimony and 2) likelihood. Different configurations and operators described in the literature are compared and evaluated: four distance methods to build the initial topology of trees, multiple rearrangement strategies, and two local search algorithms. Data sets from the related literature are used to compare our proposal with classical single and multiobjective proposals. Finally, a biological evaluation using an amino acids data set is performed. The main contributions of this paper with regard to previous publications are as follows.

1) A thorough evaluation of an NSGA-II-based strategy for the phylogenetic inference problem, considering local search strategies, and different crossover and mutation operators.
2) A characterization of different crossover, mutation and two local search strategies applied to the multiobjective phylogenetic inference problem, and its effect on the search process.
3) The proposal of a new *ad-hoc* crossover operator for the phylogenetic inference problem, which combines the parameters of the evolutionary model employed in the likelihood calculation.
4) New solutions for the literature data sets that improve the quality metrics of state of the art single and multiobjective strategies. This contribution makes the proposal a real alternative for the field.

The remainder of this paper is organized as follows. Section II introduces some concepts about phylogenetic inference and multiobjective optimization. It includes a review of the related work (Section II-D). Section III describes details of the proposal. Section V shows the results. The last section presents the main conclusions reached in this paper.

II. Phylogenetic Inference

A. Phylogenetic Trees

A phylogenetic tree represents a hypothesis about the evolutionary relations between species (or, in specific cases, among molecules themselves, e.g., protein evolution). This tree can be classified as rooted or unrooted. A rooted tree infers the existence of a common ancestor and indicates the direction of the evolution. On the contrary, an unrooted tree shows the evolutionary relationship among organisms without a common ancestor [20]. The number of possible topologies to infer rooted (nr) and unrooted (nu) trees for n species can be computed using the following equation:

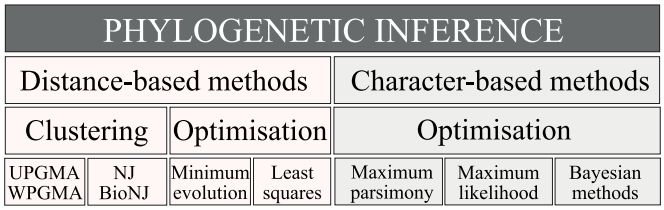

Fig. 1. Classification of the phylogenetic inference methods.

$$\mathrm{nr}(n) = \frac{(2n-3)!}{(n-2)!2^{n-2}} \tag{1}$$

$$\mathrm{nu}(n) = \frac{(2n-5)!}{(n-3)!2^{n-3}}. \tag{2}$$

Then, the search for the best rooted tree according to one criterion for only 20 species requires the evaluation of 8.2×10^{21} topologies.

B. Methods for Phylogenetic Inference

The reconstruction of phylogenetic trees can be made using distance-based and character-based methods (Fig. 1). The former infer phylogeny using a distance matrix with pairwise distances between sequences, and the latter use a set of aligned sequences and evolutionary information of the characters.

Distance-based methods can be classified as clustering and optimization methods. The first category considers greedy approaches that itself can build trees such as unweighted pair group method with arithmetic mean (UPGMA), weighted pair group method with arithmetic mean (WPGMA), neighbor joining (NJ), and bio NJ (BioNJ) [21]. The second category has two optimality criteria: 1) minimum evolution [22], [23] and 2) least-square error [24], [25]. Minimum evolution considers the shortest sum of branch lengths for choosing the best tree. In contrast, least-square error minimizes the difference between the observed pairwise distances and the distances over a phylogenetic tree.

Character-based methods use an optimization approach. The most widely used methods are maximum parsimony, maximum likelihood, and Bayesian methods [26]. Maximum parsimony considers the tree which minimizes the number of changes required to explain the input data. Maximum likelihood chooses the tree with the highest likelihood in relation to the observed data according to a specific evolutionary model [21]. Finally, the Bayesian method is based on the posterior probability, which is obtained from the likelihood and the prior probability.

Different authors have demonstrated that the maximum likelihood tree corresponds to the maximum parsimony tree, but not vice verse, in specific cases which considers symmetric-state evolutionary models with *no common mechanism*, unbounded substitution probabilities, and cases that do not include molecular clocks [27]. However, the equivalence between parsimony and likelihood by considering evolutionary models with *common mechanism* and most of the evolutionary

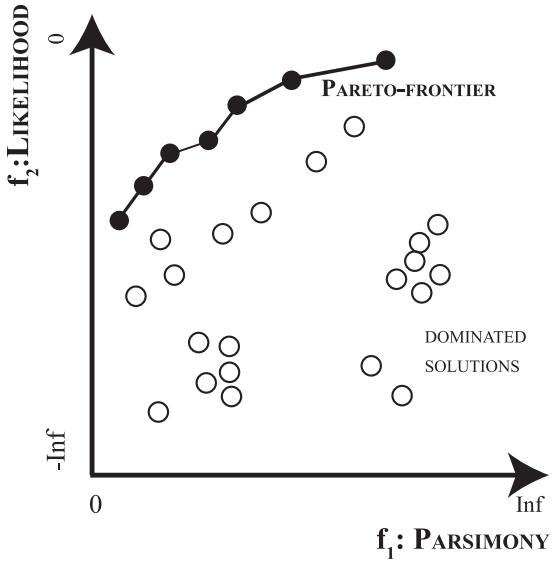

Fig. 2. Pareto-set of solutions in multiobjective phylogenetic inference.

models have not been proved. In general conditions, the parsimony and likelihood have a conflicting relationship resulting in different evolutionary hypotheses for a given data set.

C. Multiobjective Phylogenetic Inference

A single optimization problem considers the maximization (or minimization) of only one objective function. On the other hand, a multiobjective optimization problem involves multiple objectives, and they usually have some level of conflict [10]. In the phylogenetic inference context, two criteria could result in different topologies of trees for the same data set. Also, the relationships between criteria have not been well established (e.g., parsimony and maximum likelihood). In this case, the multiobjective phylogenetic inference problem can be defined as

$$\text{maximize } \vec{z} = \vec{f}(x) = (f_1(x), f_2(x)), x \in X \tag{3}$$

where x is a solution tree in the set of all possible solutions X, and $z = \vec{f}(x)$ is an objective vector point, where f_1 corresponds to the parsimony function and f_2 the likelihood function. It is important to clarify that the maximization of the parsimony involves lowering its value. The Pareto optimal solutions are those for which no other solution is better in all objectives. The points in the objective space corresponding to the Pareto-optimal are called nondominated, and they form the Pareto-frontier. In particular for this problem, we are looking for solutions that go toward the upper left of the objective space (Fig. 2).

D. Related Work

The first work applying optimization methods for phylogenetic inference used single-objective approaches. A complete review of these methods can be found in [7] and references therein. In this paper we deal with the multiobjective version of the phylogenetic inference problem.

One of the first strategies used to deal with multiobjective problems was the NSGA proposed by [28]. A second version of this algorithm (NSGA-II) was proposed by Deb *et al.* [19] which reduces its computational complexity. Also, memetic algorithms have been successfully applied to solve single and multiobjective problems in bioinformatics [29], [30] and other areas [31]–[33]. In addition, different local search strategies have been successfully integrated in multiobjective approaches [34]–[37].

In particular for the multiobjective phylogenetic inference problem, the first proposed work was developed by Poladian and Jermiin [5]. They applied an evolutionary multiobjective algorithm to optimize maximum likelihood and infer different phylogenetic trees using two biological sources in conflict: 1) mitochondrial and 2) nuclear gene information. A year later, Jayaswal *et al.* [12] applied the same method using conflicting biological sources from simian sequences to obtain different evolutionary hypotheses. The same year, Cancino and Delbem [9] proposed a multiobjective evolutionary algorithm: PhyloMOEA. It uses maximum parsimony and likelihood criteria to evaluate the evolutionary hypotheses of four data sets of nucleotide sequences. They used tree rearrangement strategies in the crossover (Lewis's operator [38]) and mutation operators [nearest neighbor interchange (NNI)].

Bio-inspired approaches have been explored for dealing with the phylogenetic inference. Coelho *et al.* [13] designed an immune-inspired multiobjective strategy to infer phylogeny using distance-based criteria: minimum evolution and least squares. Genetic operations were applied directly to distance matrices without the use of rearrangement methods. Another bioinspired approach was presented by Santander-Jiménez and Vega-Rodríguez [7]. They used a multiobjective adaptation of the artificial bee colony (MO-ABC) to maximize parsimony and likelihood using real data sets with nucleotide sequences. Their metaheuristic incorporated NNI and it was compared with NSGA-II using prune-delete-graft (PDG). In addition, their proposal was contrasted with classical single-objective optimization methods. In a previous work, they compared different mutation operators for this meta-heuristic [18]. The same author proposed a multiobjective firefly algorithm (MO-FA) for inferring phylogenetic trees according to maximum parsimony and maximum likelihood criteria [14]. They also studied the behavior of several clustering methods for this approach [15].

Parallel versions of these strategies have also been explored [39], [40]. Additionally, Santander-Jiménez and Vega-Rodríguez [17] tackled the reconstruction of phylogenetic relationships applying a parallel indicator-based evolutionary algorithm using the hypervolume metric [41]. Recently, Zambrano-Vega *et al.* [1] designed MO-phylogenetic, a software tool to infer phylogenetic trees by maximizing parsimony, and likelihood. This tool offers different mutation and crossover operators.

The design of different methods to infer phylogeny requires the use of diverse rearrangement strategies to search for optimal tree structures. Three of these operators have been widely used in the literature as mutation operators in the evolutionary algorithms [26]: 1) NNI; 2) subtree pruning

and regrafting (SPR); and 3) tree bisection and reconnection (TBR). NNI exchanges subtrees from a random internal branch to obtain a new tree. SPR selects a random subtree from a tree, removes the selected subtree and regrafts it in a random position to generate a new tree [7]. TBR combines both strategies. Other rearrangement strategies have been applied as crossover operator in genetic algorithms [42]–[45]. However, the most recent works use PDG as the crossover operator [46]. It takes a random subtree from one of the parents and inserts it in the other parent at a randomly selected insertion point, deleting duplicated species from the second tree [7]. It has been reported that crossover strategies are biased and more destructive to one parent than to the other [38], [47].

The application of multiobjective optimization strategies in the phylogenetic inference context is a current focus of research [48], mainly concentrated on the development, optimization, and evaluation of new algorithmic strategies. Despite these advances, and the clear advantage of using multiobjective approaches, the most current tools used to infer phylogenetic trees are based on single objective optimization.

III. Multiobjective Memetic Algorithm

We adapted the NSGA-II algorithm developed by Deb *et al.* [19] by integrating an *ad-hoc* local search operator and tree rearrangement strategies to tackle the phylogenetic inference problem. Algorithm 1 shows the pseudo-code of the proposal, where D corresponds to a data set with sequences in PHYLIP format, ps is the population size, cr and mr are the crossover and mutation rates, respectively, and ls corresponds to the number of local search iterations to be performed by the algorithm. The following sections describe details of the algorithm.

A. Optimality Criteria

Character-based methods use aligned sequences directly during tree inference. These methods are statistically more consistent than distance-based methods, due to the inevitable loss of evolutionary information when a sequence alignment is converted to pairwise alignment [49]. Consequently, the majority of the multiobjective optimization approaches infer phylogeny by maximizing parsimony and likelihood criteria. In our proposal, the parsimony score is calculated using Fitch's algorithm [26], and the likelihood score is obtained using Nguyen *et al.*'s [50] stochastic search algorithm. Both criteria are obtained using the *phangorn R package* [51].

B. Initial Population

A first tree T_i is built applying a distance method (UPGMA, WPGMA, NJ, or BioNJ). Its evolutionary model is estimated according to Akaike's information criterion [26]. Next, a new tree T_p is created optimizing parsimony, and its branch length is estimated using the accelerated transformation method [52]. The evolutionary model of this tree is calculated. Subsequently, a third tree T_l is built optimizing likelihood according to the evolutionary model defined for T_i. Finally, an initial population with ps individuals is built over the initial trees (T_p and T_l) using a rearrangement method (NNI, SPR, or TBR). After the initial population P is built, a second population Q is constructed using genetic operators.

Algorithm 1 MO-MA

```
1: Input: D,ps, cr, mr, ls
2: Output: A P population of trees (Pareto frontier).
                                   ▷ Initialise population
3: P ← initialise_population(D,ps)
4: while stop condition is not reached do
5:     for each p ∈ P do
                                   ▷ Genetic operations
6:         [T1, T2] ← binary_tournament_selection(P)
7:         Q[p] ← crossover(Ti, T2, cr)
8:         Q[p] ← mutation(Q[p], mr)
9:     end for
                          ▷ Update Pareto-frontier population
10: P ← non_dominated_sorting(P, Q, ps)
                                   ▷ Local search application
11: P ← local_search_strategy(P, ls)
12: end while
13: return (P)
```

C. Crossover Operators

In order to apply the crossover of solutions, two parents T_1 and T_2 are selected from P using a stochastic binary tournament. Then, the two parents are merged using rearrangement strategies. For this purpose, four strategies have implemented: 1) PDG [46]; 2) a modified version of PDG (PDGm); 3) branch exchange (BE); and 4) a CS tree method. The PDG operator was designed according to [46]. A small modification of the PDG algorithm was performed in order to build a second operator PDGm. The selection of a random subtree from one of the parents was replaced by the selection of the smallest subtree which has two or more tips (leaves). BE operator prunes a random branch from one of the parents and inserts it in the other parent at a randomly selected insertion point, deleting the duplicated species from the latter. Finally, a greedy operator (CS) was designed to search a CS tree among the parents. It applies the NNI operator on the T_1 parent until the offspring reaches a random determined distance r (Robinson–Foulds distance [53]) from both parents (in Fig. 3, we show a scheme of the four crossover operators inspired on the work of [54]). Finally, a uniform crossover operator combines the parameters of the evolutionary model employed in the likelihood calculation for each parent.

D. Mutation Operators

Tree rearrangement heuristics are used as mutation operators: NNI, SPR, and TBR. The first two were performed using the *phangorn R package*, and the third strategy was coded as a new function.

E. Local Search Strategies

The neighborhood of a population is visited using the same mutation operators described in Section III-D, as follows: the mutation operator is applied for each solution $p \in P$ to generate a p' tree. If p does not dominate p', the latter is added to the population P and a nondominated sorting

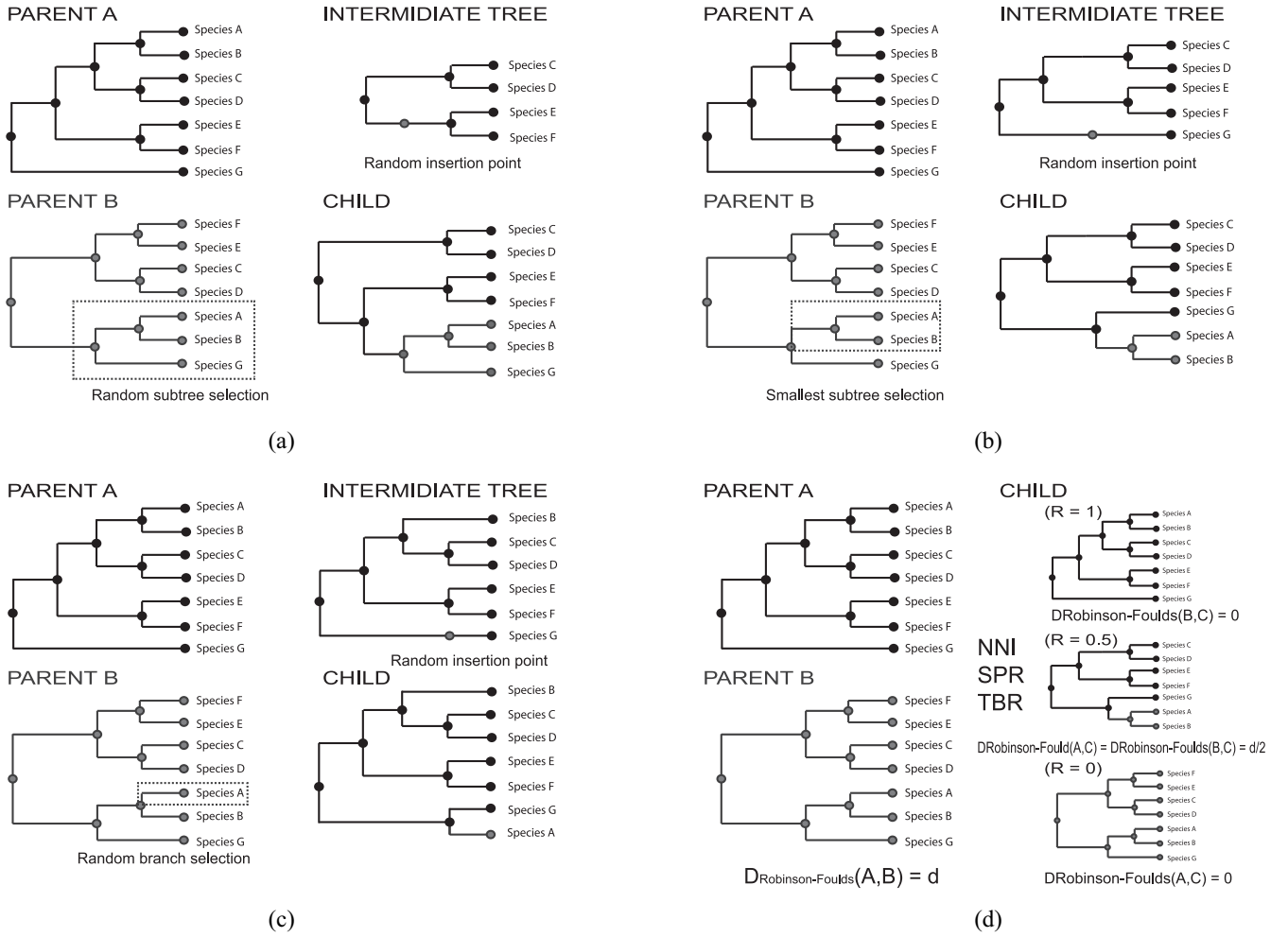

Fig. 3. Scheme of different crossover operators. (a) PDG. (b) PDGm. (c) BE. (d) CS.

algorithm with crowding distance is applied to discard a solution. This operation is applied using: 1) a Pareto local search algorithm [36], [37] and 2) a simulated annealing algorithm [55]. The former runs for a fixed number ls of iterations.

IV. Experimental Design

In order to identify the configuration that maximizes the performance of the multiobjective memetic algorithm (MO-MA), we tested and characterized different tree rearrangement strategies as genetic operators and then they were integrated to the structure of MO-MA to evaluate multiple global configurations. The best global configuration found was tested against other literature strategies [7], [9], [11], [13]–[18], [39], [40] and it was also applied to study an amino acid data set of biological interest.

A. Configuration of the Algorithm

1) Crossover Operator Evaluation: We tested two crossover operator conditions: 1) the bias associated with the production of descendants which include characteristics of both parents (balance condition) and 2) the ability to visit the search space. To test the first condition, we applied a 1000 crossover operation using the rearrangement methods over the initial topologies. Then, the distance between descendant and each couple of parents was measured using the Robinson–Foulds metric, normalized by the distance between parents.

In order to measure the capacity of each rearrangement method to visit the search space and obtain new solutions, we also compute the average normalized distance between the descendants and both parents. The goal is to know how similar are the parents and the offspring generated after the crossover operator. The data sets selected and the statistical comparison methods applied are the same as those presented in Section IV-A3.

2) Mutation Operator Evaluation: We also tested the ability of each mutation operator to visit the search space using the Robinson–Foulds metric. One thousand mutations were generated over the initial topologies, and the same methodology employed in the crossover analysis was applied.

3) Global Configuration: We tested the memetic algorithm using different configurations: four distance methods to generate the initial topologies (UPGMA, WPGMA, NJ, and BioNJ), five crossover alternatives (PDG, PDGm, BE, CS, and no crossover operator), four mutation strategies (NNI, SPR, TBR, and no mutation operator), and three local search strategies (greedy strategy, simulated annealing, and no local search strategy). The number of possible combinations is 240

TABLE I
DATA SETS AND CORRESPONDING REFERENCES APPLIED IN TESTS

Data sets	Sequences	#Sequences	#Nucleotides	References
primates_14	Mitoch. D-loop and adjacent 3rd positions for apes	14	232	[64]
rbcL_55	rbcL gene	55	1314	[1, 7, 11, 13, 14]
HIV2_72	HIV2 from (HIV Sequence Database)	72	828	[7, 14]
membracidae_81	EG-1a and 28S rDNA	81	3321	[7, 14]
ureases_126	ureases enzyme - amino acid	126	609	[63]
mtDNA_186	Human mitochondrial DNA	186	16608	[7, 11, 13, 14]
HIV1_192	HIV1 from (HIV Sequence Database)	192	817	[7, 14]
RDPII_218	Prokaryotic RNA	218	4182	[1, 7, 11, 13, 14]
ZILLA_500	rbcL plastid gene	500	759	[7, 11, 13, 14]

$(4 \times 5 \times 4 \times 3)$. Each configuration was ran 30 times, resulting in 7200 executions for each data set. The data sets used are primates_14, rbcL_55, and HIV1_192.

In order to evaluate the quality of solutions, we computed the hypervolume metric of the Pareto-frontier. We applied the Kruskal–Wallis test by ranks to compare the statistically significant difference in performance between configurations. The *post hoc* analysis was performed using the Dunn test with a significance level of 1%. Furthermore, the coverage metric [56] and the percentage representativeness in the global Pareto-frontier were calculated using the Pareto-set of solutions corresponding to the median hypervolume achieved by each configuration (representative solution). This method has been applied in previous works [7]. The global Pareto-frontier considers all the nondominated solutions from all the configurations.

To perform the evaluation and the comparison with other strategies, we configured MO-MA with parameters previously recommended by the literature for the population size (ps) and the genetic parameters (cr and mr) [7], [11]. The parameters for both local search strategies were experimentally defined maximizing the hypervolume metric according to the strategy proposed in [14] and [39] (see details in the supplementary material, Appendix S1-A).

B. Performance Evaluation

1) Comparison With Single-Objective Optimization Approaches: In order to show the benefit of dealing with the multiobjective version of the problem over the single objective version, we compared MO-MA with other widely used single-objective optimization tools such as PHYML (Sankoff parsimony), DNAPARS, RAxML [57], and MEGA [58] (parsimony and likelihood tree). Each algorithm was executed 30 times for each data set. The difficulties in comparing single and multiobjective optimization evolutionary algorithms have been discussed in [59]. We adjusted the comparison according to the hypervolume contribution of each nondominated solution, dividing their hypervolume by the cumulative sum of the hypervolume of all the Pareto-set of solutions. The coverage and the representativeness in the global Pareto-frontier were calculated using the method proposed in Section IV-A3, and the parameterization of each tool is detailed in the supplementary material (Appendix S1-B).

2) Comparison With Multiobjective Optimization Approaches: In order to show the performance of the proposal, MO-MA was compared with other multiobjective methods proposed in the literature: NSGA-II [7], PhyloMOEA [9], MO-ABC [7], MO-FA [14], MO-Phyl [39], and Mophylogenetics [1]. Furthermore, we implemented an NSGA-II which includes the crossover operator combining the parameters from the evolutionary model (NSGA-II EM). This algorithm and MO-MA were executed 30 times for each data set. The median of hypervolume was calculated, and the representative set of solutions was used to perform the comparison with other tools. The representative Pareto-frontiers obtained by the other proposals were extracted from the plots published for each method using the WebPlotDigitizer tool 3.11 [60], [61]. To perform a fair comparison, the number of generations of MO-MA was limited to the time required by NSGA-II to iterate 100 generations (supplementary material, Appendix S1-C).

3) Study of Amino Acids Data Set: We design this test to evaluate: 1) the ability of our proposal to infer evolutionary hypotheses using amino acid data sets, since all the current-state proposals have been tested using nucleotide data sets and 2) to evaluate the biological meaning of the solutions.

Ureases are of special interest, since it is very hard to propose an evolutionary history of these multifunctional enzymes, due to their varied structural organization [62]. An accepted evolutionary hypothesis for these taxa has been previously

TABLE II
ROBINSON–FOULDS DISTANCE BETWEEN DESCENDANTS AND EACH PARENT (P1, P2) OBTAINED USING THE CROSSOVER OPERATORS (MEDIAN AND STANDARD DEVIATION)

Crossover	Parents	rbcL_55	HIV1_192	ZILLA_500
PDG	P1	1.7 (0.4)	1.0 (0.1)	1.1 (0.0)
	P2	1.0 (0.3)	0.2 (0.1)	0.2 (0.0)
PDGM	P1	2.5 (0.2)	1.0 (0.0)	1.1 (0.0)
	P2	1.8 (0.5)	0.2 (0.0)	0.2 (0.0)
CS	P1	**0.1 (1.9)**	**0.1 (0.7)**	**3.3 (1.7)**
	P2	**1.1 (1.3)**	**1.0 (0.2)**	**1.4 (1.2)**
BE	P1	1.9 (0.3)	1.0 (0.0)	1.0 (0.0)
	P2	1.1 (0.4)	0.1 (0.0)	0.1 (0.0)

TABLE III
AVERAGE ROBINSON–FOULDS DISTANCE BETWEEN OFFSPRING AND PARENTS AFTER THE APPLICATION OF DIFFERENT GENETIC OPERATORS

Crossover	rbcL_55	HIV1_192	ZILLA_500
PDG	1.6 (0.4)	0.5 (0.5)	0.6 (0.5)
PDGM	2.2 (0.4)	0.6 (0.4)	0.7 (0.5)
CS	**3.6 (1.4)**	**1.3 (0.2)**	**1.0 (1.7)**
BE	1.6 (0.4)	0.6 (0.5)	0.6 (0.5)
Mutation	**rbcL_55**	**HIV1_192**	**ZILLA_500**
NNI	2.0 (0.0)	2.0 (0.0)	2.0 (0.0)
SPR	**18.0 (6.0)**	**26.1 (8.4)**	**37.7 (10.3)**
TBR	**16.8 (5.3)**	**30.9 (8.1)**	**40.3 (9.3)**

TABLE IV
TOP FIVE BEST AND WORST MEMETIC ALGORITHM CONFIGURATIONS ACCORDING TO THE HYPERVOLUME METRIC (*hyp.*)—MEDIAN AND STANDARD DEVIATION. THE COVERAGE METRIC (*cov.*) REPRESENTS THE PERCENTAGE OF NONDOMINATED SOLUTIONS. THE REPRESENTATIVENESS IN THE GLOBAL PARETO-FRONTIER (*% Par. Front.*) CORRESPONDS TO THE RATIO BETWEEN THE NONDOMINATED SOLUTIONS OBTAINED FOR A SPECIFIC METHOD, AND THE TOTAL NUMBER OF SOLUTIONS IN THE PARETO-FRONTIER. THE NJ-PDG-NNI-G CONFIGURATION HAS A GOOD PERFORMANCE IN ALL THE STUDIED DATA SETS (GRAY COLOR). (*ini: Initialization,* cr*: Crossover, mut: Mutation, and* ls*: Local Search*)

Data Set	Best configuration						% Par.	Worst configuration						% Par.
	ini	cr	mut	ls	hyper.	cov.	Front	ini	cr	mut	ls	hyper.	cov.	Front
rbcL_55	NJ	PDG	NNI	G	3.67(0.1)	0%	31%	WPGMA	NONE	TBR	-	0.60(1.3)	50%	0%
	UPGMA	CS	NNI	G	3.63(0.1)	0%	14%	WPGMA	BE	TBR	SA	0.33(1.0)	50%	0%
	BIONJ	CS	NNI	G	3.52(0.3)	0%	2%	BIONJ	PDGM	SPR	-	0.30(0.9)	50%	0%
	NJ	NONE	NNI	SA	3.45(0.2)	5%	1%	WPGMA	-	SPR	-	0.30(0.9)	100%	0%
	BIONJ	BE	NNI	-	3.34(0.2)	0%	1%	WPGMA	PDG	SPR	-	0.30(0.9)	100%	0%
HIV1_192	NJ	PDG	NNI	G	3.09(0.7)	0%	25%	WPGMA	BE	TBR	G	0.00(0.0)	100%	0%
	UPGMA	PDG	NNI	SA	3.00(0.0)	0%	10%	WPGMA	BE	TBR	SA	0.00(0.0)	100%	0%
	UPGMA	CS	NNI	-	2.72(0.9)	0%	2%	WPGMA	-	TBR	SA	0.00(0.0)	100%	0%
	NJ	CS	NNI	SA	2.58(1.1)	0%	1%	WPGMA	PDGM	TBR	SA	0.00(0.0)	100%	0%
	WPGMA	PDG	NNI	SA	2.43(1.3)	100%	0%	WPGMA	PDG	TBR	G	0.00(0.0)	100%	0%
primates_14	WPGMA	PDG	NNI	G	3.39(0.0)	1%	5%	BioNJ	PDG	TBR	SA	3.08(0.1)	98%	1%
	UPGMA	PDG	NNI	G	3.39(0.0)	2%	20%	BioNJ	PDGM	TBR	G	3.06(0.1)	98%	1%
	NJ	PDG	NNI	G	3.39(0.0)	0%	55%	NJ	PDG	TBR	G	3.00(0.0)	98%	1%
	BioNJ	CS	NNI	G	3.37(0.1)	98%	1%	NJ	PDG	TBR	SA	3.00(0.0)	98%	2%
	BioNJ	CS	TBR	SA	3.27(0.1)	98%	1%	WPGMA	PDGM	TBR	SA	3.00(0.0)	65%	0%

TABLE V
HYPERVOLUME CONTRIBUTION, COVERAGE METRICS, AND REPRESENTATIVENESS IN THE GLOBAL PARETO-FRONTIER FOR MO-MA AND THE SINGLE-OBJECTIVE OPTIMIZATION APPROACHES. THE BEST VALUES HAVE BEEN HIGHLIGHTED IN BOLD FOR EACH DATA SET. IF MANY STRATEGIES HAVE BOLDED VALUES FOR THE SAME DATA SET, THEIR DIFFERENCE IS NOT STATISTICALLY SIGNIFICANT

Hypervolume of each point in Pareto Frontier						
Data sets	MO-MA	DNAPARS	MEGAlik	MEGApar	PHYML	RAxML
primates_14	**33% (1%)**	0%	0%	0%	0%	0%
rbcL_55	**14% (0%)**	0%	0%	0%	0%	0%
HIV2_72	**18% (1%)**	0%	0%	0%	0%	11%
membracidae1_81	**25% (0%)**	0%	0%	0%	0%	0%
mtDNA_186	0% (0%)	0%	0%	0%	0%	**100%**
HIV1_192	**34% (0%)**	0%	0%	0%	0%	32%
RDPII_218	**9% (0%)**	8%	0%	0%	0%	**9%**
ZILLA_500	**13% (0%)**	12%	11%	0%	**13%**	0%
Coverage metric						
Data sets	MO-MA	DNAPARS	MEGAlik	MEGApar	PHYML	RAxML
primates_14	**0%**	100%	100%	100%	100%	100%
rbcL_55	**0%**	100%	100%	100%	100%	100%
HIV2_72	**0%**	100%	100%	100%	100%	**0%**
membracidae1_81	**0%**	100%	100%	100%	100%	100%
mtDNA_186	100%	100%	100%	100%	100%	**0%**
HIV1_192	**0%**	100%	100%	100%	100%	**0%**
RDPII_218	38%	**0%**	100%	100%	100%	**0%**
ZILLA_500	50%	**0%**	**0%**	100%	**0%**	100%
% Solutions in Pareto-frontier						
Data sets	MO-MA	DNAPARS	MEGAlik	MEGApar	PHYML	RAxML
primates_14	**75%**	0%	0%	0%	0%	0%
rbcL_55	**100%**	0%	0%	0%	0%	0%
HIV2_72	**83%**	0%	0%	0%	0%	17%
membracidae1_81	**100%**	0%	0%	0%	0%	0%
mtDNA_186	0%	0%	0%	0%	0%	**100%**
HIV1_192	**67%**	0%	0%	0%	0%	33%
RDPII_218	**72%**	7%	0%	0%	0%	7%
ZILLA_500	**64%**	**12%**	**12%**	0%	**12%**	0%

inferred by using MEGA [63]. The evolutionary hypotheses were compared by applying the Shimodaira–Hasegawa test [9]. It calculates the statistically significant difference between the most likelihood trees obtained by MO-MA and the accepted topology proposed in [63].

V. EXPERIMENTAL RESULTS

The algorithms included in this paper and the statistics test were implemented using *R* version 3.2.3 and RStudio version 0.99.491. The data sets were taken from the related literature, and Table I shows details of the sequences and their corresponding sources.[1]

[1]Code and data sets are available from the online resource (http://bioinformatic.diinf.usach.cl/phylogeny/phylogeny.html).

A. Configuration of the Algorithm

1) Crossover Operator Evaluation: The difference of the Robinson–Foulds distance between descendants and each parent for the crossover operators is shown in Table II. For crossovers PDG, PDGm, and BE, the difference between the offspring and their parents is statistically significant, which means that this operator generates offspring that have more similarity to one of the parents. Then, these operators show a bias when generating the offspring (unbalanced condition). In contrast, the CS operator presents a balanced condition, since the random variable that controls the similarity of the descendants with parents reduces the bias of the operator.

In the second experiment, we measure the average distance between offspring and their parents (see Table III). For

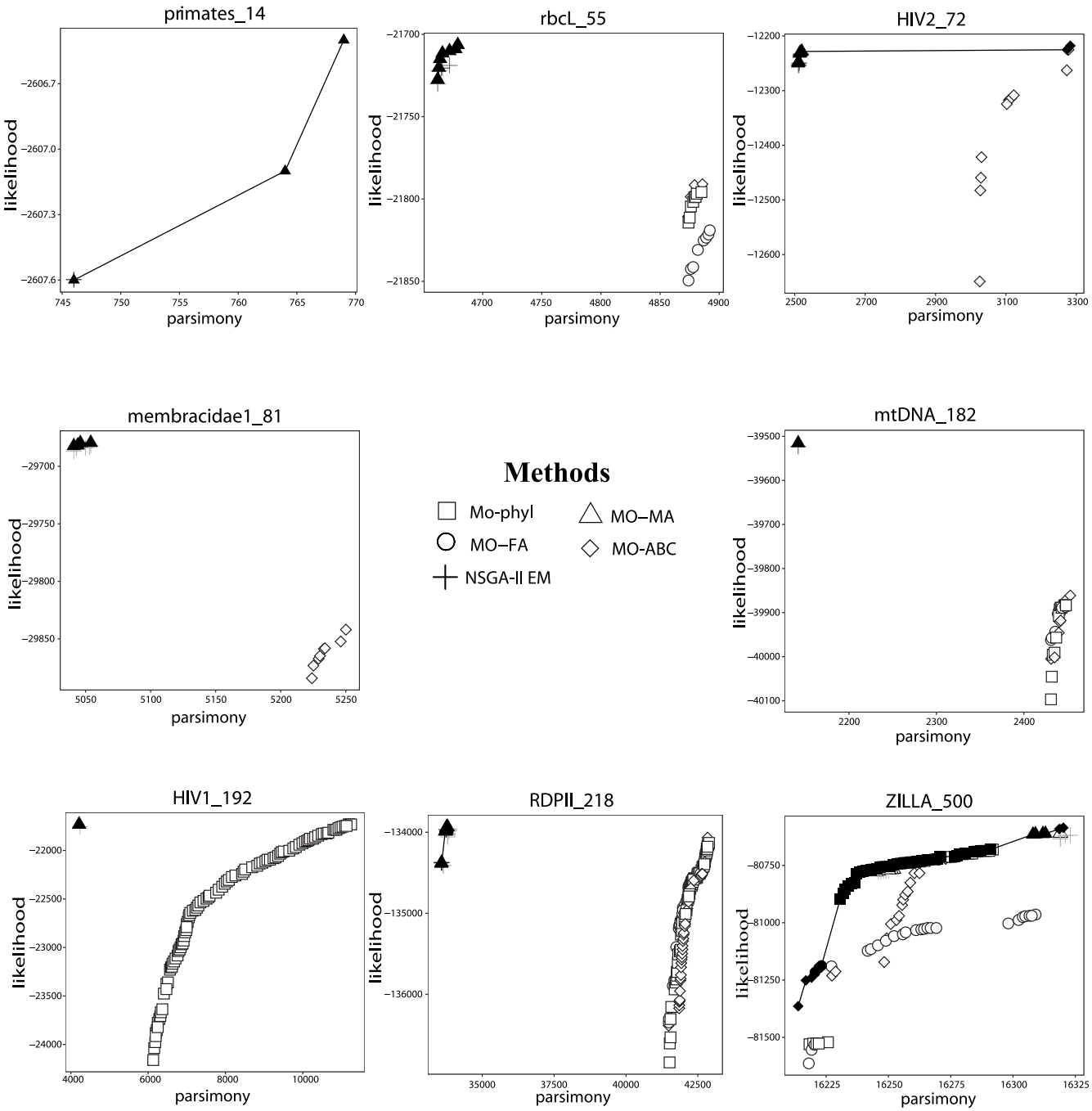

Fig. 4. Representative Pareto-frontier from MO-MA and other multiobjective optimization approaches. The approximated global Pareto-frontier is shown in black. Solutions from PhyloMOEA, MO-Phylogenetics, and NSGA-II are not represented because their fitness values affect the visualization of the plots.

crossovers PDG, PDGm, and BE, the results obtained are not statistically significant, which means that they generate offspring that are similar to their parents. However, CS has a greater average distance, and the analysis shows that this difference is statistically significant. This result indicates that CS is able to generate solutions that are different from their parents, allowing to visit farther regions of the solution space.

Details can be seen in the supplementary material (Appendix S1-D and Appendix S1-E).

2) Mutation Operator Evaluation: Table III shows the Robinson–Foulds distance after the application of different mutation operators. Naturally, descendants from NNI operators have a lower distance in relation to SPR and TBR. By definition, the distance of NNI descendants always will be equal to 2. A statistically significant difference between SPR and TBR was not found. Additional details can be seen in the supplementary material (Appendix S1-E).

3) Global Configuration: A total of 240 configurations of the MO-MA on three data sets were tested (rbcL_55, HIV1_192 and primates_14). Table IV shows the five best and worst configurations according to the hypervolume metric, the coverage, and the representativeness in the Pareto-frontier.

In Table IV, it is seen that the configuration that uses population operators NJ, PDG, NNI, and the greedy local

TABLE VI
HYPERVOLUME, COVERAGE METRIC, AND REPRESENTATIVENESS OF THE GENERAL PARETO-FRONTIER FOR MO-MA AND OTHER MULTIOBJECTIVE OPTIMIZATION APPROACHES. THE BEST VALUES HAVE BEEN HIGHLIGHTED IN BOLD FOR EACH DATA SET. IF MANY STRATEGIES HAVE BOLDED VALUES FOR THE SAME DATA SET, THEIR DIFFERENCE IS NOT STATISTICALLY SIGNIFICANT

Hypervolume metric								
Data sets	MO-MA	NSGA-II EM	MO-ABC	MO-FA	MO-Phyl	MO-Phylo	NSGA-II	PhyloMOEA
primates_14	**2.20 (0.33)**	**2.14 (0.05)**	-	-	-	-	-	-
rbcL_55	**4.00 (0.00)**	3.99 (0.00)	2.13	2.11	2.12	2.13	2.12	1.10
HIV2_72	**3.96 (0.02)**	**3.96 (0.02)**	2.59	-	-	-	2.58	-
membracidae1_81	**3.99 (0.01)**	**3.97 (0.00)**	1.68	-	-	-	1.57	-
mtDNA_186	**4.00 (0.00)**	**4.00 (0.00)**	1.90	1.89	1.89	-	1.83	1.13
HIV1_192	**4.00 (0.00)**	3.99 (0.00)	-	-	3.27	-	-	-
RDPII_218	**4.00 (0.00)**	3.99 (0.00)	2.29	2.28	2.28	2.27	2.26	1.24
ZILLA_500	3.46 (0.01)	3.45 (0.01)	**3.96**	3.80	3.88	-	3.54	2.03
Coverage metric								
Data sets	MO-MA	NSGA-II EM	MO-ABC	MO-FA	MO-Phyl	MO-Phylo	NSGA-II	PhyloMOEA
primates_14	**0%**	**0%**	-	-	-	-	-	-
rbcL_55	**0%**	100%	100%	100%	100%	100%	100%	100%
HIV2_72	**0%**	100%	86%	-	-	-	100%	-
membracidae1_81	**0%**	100%	100%	-	-	-	100%	-
mtDNA_186	**0%**	100%	100%	100%	100%	100%	100%	100%
HIV1_192	**0%**	100%	-	-	100%	-	-	-
RDPII_218	**0%**	**0%**	100%	100%	100%	100%	100%	100%
ZILLA_500	60%	100%	73%	89%	**49%**	-	100%	100%
% Solutions in global Pareto-frontier								
Data sets	MO-MA	NSGA-II EM	MO-ABC	MO-FA	MO-Phyl	MO-Phylo	NSGA-II	PhyloMOEA
primates_14	**75%**	25%	-	-	-	-	-	-
rbcL_55	**100%**	0%	0%	0%	0%	0%	0%	0%
HIV2_72	**71%**	0%	29%	-	-	-	0%	-
membracidae1_81	**100%**	0%	0%	-	-	-	0%	-
mtDNA_186	**100%**	0%	0%	0%	0%	-	0%	0%
HIV1_192	**100%**	0%	-	-	0%	-	-	-
RDPII_218	**84%**	16%	0%	0%	0%	0%	0%	0%
ZILLA_500	9%	0%	14%	7%	**70%**	-	0%	0%

search algorithm (G) shows the highest hypervolume metric for the three data sets. In the case of the rbcL_55 data set, this configuration reaches better hypervolume values than all the other configurations, which are significant in 54% of the cases compared to the other configurations. On the other two data sets, although the selected configuration reaches better hypervolume values, they are statistically significant in 6% and 1% of the comparisons. This can be accounted for by the high standard deviation reached on the HIV1_192 data set, and also the small size of the primates_14 data set can lead to most of the configurations converging to similar solutions.

According to the coverage metric, the solutions found by the NJ-PDG-NNI-G configuration are nondominated by the other configurations, and also they have the largest number of solutions (31%, 25%, and 55%, respectively) in the global Pareto-frontier. The latter represents the contribution of the selected configuration to the construction of the global solution.

In the case of the configurations with the lowest hypervolume metric, it is not possible to identify a common configuration considering all data sets. However, WPGMA is part of the worst configurations. Based on these results, the NJ-PDG-NNI-G configuration was selected as the base structure for MO-MA, to perform the comparison with the state of the art alternatives.

B. Performance Evaluation

1) Single-Objective Optimization Performance Evaluation: Table V shows the hypervolume contribution of each solution, coverage metric and representativeness of the global Pareto-frontier for MO-MA and the single-optimization tools. Parsimony and likelihood scores were normalized between 0 and 1. To compute the hypervolume metric we used a reference point of (2, 2). Values in bold represents the highest hypervolume contribution that are statistically significant compared to the other methods.

MO-MA has the highest hypervolume contribution in seven out of eight data sets (Table VI), and these values are statistically significant compared to the single objective methods. The coverage metric on the same data sets shows that MO-MA produced a Pareto-frontier with solutions that are nondominated by the single objective solutions, except on data sets RDPII_218 and ZILLA_500, where they have 38% and 50% of solutions dominated by single objective methods, respectively. Regarding the number of solutions in the global Pareto Frontier, our MO-MA contributes at least with 64% of the solutions, with an average of 80%.

A particular case corresponds to the data set mtDNA_186, for which the best performing tool is RAxML. This tool reaches the best values in three evaluated metrics. It is also interesting to see in the supplementary material (Appendix S1-F), that the proposed MO-MA produced a Pareto-frontier with only one solution. Both solutions reach similar levels of parsimony, but the likelihood of the solution produced by RAxML is better than any other solution.

2) Multiobjective Optimization Performance Evaluation: The comparison of multiobjective optimization approaches is shown in Table V. In order to compute the hypervolume metric, the parsimony and likelihood scores were normalized between 0 and 1 for each data set. The reference point was defined as (2, 2).

MO-MA has the highest hypervolume for seven out of eight data sets (Fig. 4). When we compared MO-MA with NSGA-II EM, the results were statistically significant in only 50% of

the cases. For the rest of the algorithms, the presented MO-MA reaches solutions with better values that are statistically significant on all data sets. The only exception is MO-ABC, which gets better values on the ZILLA_500 data set.

The coverage metric demonstrates that solutions obtained by MO-MA are nondominated by solutions from other methods in all the data sets except ZILLA_500. In this case, MO-ABC, MO-FA, and MO-Phyl also have nondominated solutions. Also, MO-MA is the method that most contributes to the global Pareto-frontier for all data sets. However, when the ZILLA_500 data set is considered, Mo-Phyl and MO-ABC also have solutions which conform part of the general Pareto-frontier. These trees minimize the parsimony score in relation to MO-MA.

These results show that the solutions found by our algorithm are new solutions, and also contribute on the average with more than 90% of the global Pareto-frontier on seven data sets. The only exception is on data set ZILLA_500, on which algorithms MO_Phyl and MO_ABC have a greater contribution to the global Pareto frontier.

3) Study of Amino Acids Data Set: The representative Pareto-frontier obtained by the MO-MA is shown in the supplementary material (Appendix S1-G). It considers three points in the objective space and five different trees (A, B, C, D, and F). Trees with equal fitness values resulted with different topologies, such as (A, D) and (C, E). The highest Robinson–Foulds distance was 16 editions. This value corresponds to the distance between the most parsimonious tree (B), and one of the trees with the highest likelihood (A). This difference demonstrates the conflict between trees obtained using parsimony and likelihood criteria.

The Shimodaira–Hasegawa test (*ape package in R*) did not report significant differences between the topology presented in [63] and the topologies from the most likelihood trees obtained in the different executions using MO-MA. It demonstrates that our MO-MA is able to produce solutions that not only have a good performances in terms of numerical scores, but they are also biologically sound, finding solutions with accepted evolutionary hypothesis using amino acid sequences.

VI. Conclusion

In this paper we propose a new approach for solving the phylogenetic inference problem based on multiobjective optimization and evolutionary techniques. The final proposal was reached after a thorough evaluation of the different operators during the design. Overall, the trees obtained with our MO-MA have higher fitness values in relation to proposals from the literature for the majority of the data sets taken from the literature.

Based on the results shown in the previous section, it was not possible to identify a single best algorithm configuration with a highest statistically significant level of hypervolume metric for all data sets. However, the NJ-PDG-NNI-G configuration gives a structure with the highest value of hypervolume metric value, the best coverage score and solutions which represent the greater part of the global Pareto-frontier.

Also, individual features could be identified. For example, the NNI mutation operator is part of the structures with the best performance, and the use of WPGMA as a method to build initial topologies generates results with a poor multiobjective metrics for all data sets.

The balance condition for the crossover operator is important, since the selection method applied for the crossover operator influences the offspring features. When a nonstochastic selection method is used with an unbalanced crossover operator, the descendants in each generation will receive characteristics from only one of the parents, increasing the risk of stagnation in local minima. On the other hand, in an advanced generation of the heuristic, this configuration could ensure the conservation of the general features of the population.

The CS operator is part of the best configuration for the three studied data sets. Furthermore, when this operator was studied separately, it was balanced, receiving equally the features from both parents, according to the Robinson–Foulds distance. Furthermore, it has the highest distance between descendants and parents in relation to other crossover operators. These results allow inferring that this operator can be used to quickly cover the search space in the early application of a meta-heuristic. Besides, when it is necessary to keep the features of a population and at the same time minimize the risk of stagnation in local minima, other operators like PDG, PDGm, and BE could be used.

In the case of tree rearrangement for the mutation stage, SPR and TBR have a great Robinson–Foulds distance with NNI. However, no statistically significant difference was found between the first two operators. The results suggest that the use of TBR does not have advantages in relation to SPR. This consideration is important because TBR requires more movements than SPR to perform an operation [65] resulting more computationally expensive.

When the MO-MA with the NJ-PDG-NNI-G configuration was compared with single and multiobjective optimization approaches proposed in the literature, it shows better performance than literature methods according to the hypervolume metric for most of the data sets. In the other cases, the MO-MA inferred trees which are not covered by the other methods and represent a different section of the global Pareto-frontier. It means that we are able to provide new solutions to the problem. In four data sets, no statistically significant difference was reported considering the hypervolume metric when MO-MA and NSGA-II EM were compared. Considering these results and the comparison with NSGA-II without the crossover operator applied to the evolutionary model, it is possible to infer that this crossover operator helps to maximize considerably the likelihood score, improving the multiobjective metrics. As expected, the local search operator performs a refinement of the solutions, turning MO-MA in a competitive method in relation to the other proposals.

In relation to the study of the amino acids data set, no statistically significant differences were reported using the Shimodaira–Hasegawa test. It shows that MO-MA has a competitive performance based not only on the algorithmic perspective, but also considering a biological meaning, finding an accepted evolutionary hypothesis.

Although the results shown in this paper are promising, there still are important issues to improve in the algorithm, such as the study of the local search stage considering different

strategies (operator, timing, stop conditions, neighborhood definition, among others). Also, the different techniques to reduce the search space using prior knowledge can be explored. This idea has been applied successfully in other areas such as structural bioinformatics [66]. Furthermore, different metrics (spread, coverage, or hypervolume), can be included as objective functions, improving the quality of the Pareto-frontier and increasing the speed of the convergence. To generate a future applicable tool, different decision-making techniques for phylogeny inference must be explored.

REFERENCES

[1] C. Zambrano-Vega, A. J. Nebro, and J. F. Aldana-Montes, "MO-phylogenetics: A phylogenetic inference software tool with multi-objective evolutionary metaheuristics," *Methods Ecol. Evol.*, vol. 7, no. 7, pp. 800–805, 2016.

[2] M. Gupta and S. Singh, "A novel genetic algorithm based approach for optimization of distance matrix for phylogenetic tree construction," *Int. J. Comput. Appl.*, vol. 52, no. 9, pp. 14–18, 2012.

[3] B. Felix, "Phylogenetics: Tracing the evolutionary legacy of organisms, metastatic clones, bioactive compounds and languages," *J. Phylogenet. Evol. Biol.*, vol. 3, no. 2, 2015, Art. no. 1000e112.

[4] C. Hinchliff *et al.*, "Synthesis of phylogeny and taxonomy into a comprehensive tree of life," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 41, pp. 12764–12769, 2015.

[5] L. Poladian and L. S. Jermiin, "Multi-objective evolutionary algorithms and phylogenetic inference with multiple data sets," *Soft Comput.*, vol. 10, no. 4, pp. 359–368, 2006.

[6] O. Gascuel, "On the optimization principle in phylogenetic analysis and the minimum-evolution criterion," *Mol. Biol. Evol.*, vol. 17, no. 3, pp. 401–405, 2000.

[7] S. Santander-Jiménez and M. A. Vega-Rodríguez, "Applying a multiobjective metaheuristic inspired by honey bees to phylogenetic inference," *Biosystems*, vol. 114, no. 1, pp. 39–55, 2013.

[8] D. L. Swofford *et al.*, "Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods," *Syst. Biol.*, vol. 50, no. 4, pp. 525–539, 2001.

[9] W. Cancino and A. C. B. Delbem, "A multi-objective evolutionary approach for phylogenetic inference," in *Proc. Int. Conf. Evol. Multi Criterion Optim.*, vol. 4403, 2007, pp. 428–442. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-70928-2_34

[10] J. Handl, D. B. Kell, and J. Knowles, "Multiobjective optimization in bioinformatics and computational biology," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 4, no. 2, pp. 279–292, Apr./Jun. 2007.

[11] G. P. Coelho, F. J. V. Zuben, and A. E. A. da Silva, "A multiobjective approach to phylogenetic trees: Selecting the most promising solutions from the Pareto front," in *Proc. IEEE Int. Conf. Intell. Syst. Design Appl.*, 2007, pp. 837–842.

[12] V. Jayaswal, L. Poladian, and L. S. Jermiin, "Single- and multi-objective phylogenetic analysis of primate evolution using a genetic algorithm," in *Proc. IEEE Int. Conf. Evol. Comput. (CEC)*, vol. 1, 2007, pp. 4146–4153.

[13] G. P. Coelho, A. E. A. da Silva, and F. J. Von Zuben, "An immune-inspired multi-objective approach to the reconstruction of phylogenetic trees," *Neural Comput. Appl.*, vol. 19, no. 8, pp. 1103–1132, 2010.

[14] S. Santander-Jiménez and M. A. Vega-Rodríguez, "A multiobjective proposal based on the firefly algorithm for inferring phylogenies," in *Proc. Eur. Conf. Evol. Comput. Mach. Learn. Data Min. Bioinformat. (EvoBIO)*, 2013, pp. 141–152.

[15] S. Santander-Jiménez and M. A. Vega-Rodríguez, "A comparative study on distance methods applied to a multiobjective firefly algorithm for phylogenetic inference," in *Proc. Annu. Conf. Companion Genet. Evol. Comput. (GECCO)*, 2013, pp. 1587–1594.

[16] S. Santander-Jiménez and M. A. Vega-Rodríguez, "Inferring multiobjective phylogenetic hypotheses by using a parallel indicator-based evolutionary algorithm," in *Theory and Practice of Natural Computing (TPNC)*. Cham, Switzerland: Springer, 2014, pp. 205–217.

[17] S. Santander-Jiménez and M. A. Vega-Rodríguez, "Performance evaluation of dominance-based and indicator-based multiobjective approaches for phylogenetic inference," *J. Inf. Sci.*, vol. 330, pp. 293–314, Feb. 2016.

[18] S. Santander-Jiménez, M. A. Vega-Rodríguez, J. A. Gómez-Pulido, and J. M. Sánchez-Pérez, "Comparing different operators and models to improve a multiobjective artificial bee colony algorithm for inferring phylogenies," in *Theory and Practice of Natural Computing (TPNC)*, vol. 7505. Heidelberg, Germany: Springer, 2012, pp. 187–200. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-33860-1_16

[19] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[20] T. Strachan and A. P. Read, *Human Molecular Genetics*, vol. 1, 4th ed. New York, NY, USA: Garland Sci., 2011.

[21] A. De Bruyn, D. Martin, and P. Lefeuvre, "Phylogenetic reconstruction methods: An overview," in *Molecular Plant Taxonomy: Methods and Protocols*, vol. 1115. Totowa, NJ, USA: Humana Press, 2014, pp. 257–277. [Online]. Available: https://link.springer.com/protocol/10.1007%2F978-1-62703-767-9_13

[22] K. K. Kidd and L. A. Sgaramella-Zonta, "Phylogenetic analysis: Concepts and methods," *Amer. J. Human Genet.*, vol. 23, no. 3, pp. 235–252, 1971.

[23] A. Rzhetsky and M. Nei, "Theoretical foundation of the minimum-evolution method of phylogenetic inference," *Mol. Biol. Evol.*, vol. 10, no. 5, pp. 1073–1095, 1993.

[24] L. L. Cavalli-Sforza and A. W. F. Edwards, "Phylogenetic analysis. models and estimation procedures," *Amer. J. Human Genet.*, vol. 19, no. 3, pp. 233–257, 1967.

[25] W. M. Fitch and E. Margoliash, "Construction of phylogenetic trees," *Science*, vol. 155, no. 3760, pp. 279–284, 1967.

[26] J. Felsenstein, *Inferring Phylogenies*, vol. 2. Sunderland, MA, USA: Sinauer Assoc. Inc., 2004.

[27] M. Fischer and B. Thatte, "Revisiting an equivalence between maximum parsimony and maximum likelihood methods in phylogenetics," *Bull. Math. Biol.*, vol. 72, no. 1, pp. 208–220, 2010.

[28] N. Srinivas and K. Deb, "Muiltiobjective optimization using nondominated sorting in genetic algorithms," *Evol. Comput.*, vol. 2, no. 2, pp. 221–248, Sep. 1994.

[29] C. Clark and J. Kalita, "A multiobjective memetic algorithm for PPI network alignment," *Bioinformatics*, vol. 31, no. 12, pp. 1988–1998, 2015.

[30] Á. Rubio-Largo, M. A. Vega-Rodríguez, and D. L. González-Álvarez, "A hybrid multiobjective memetic metaheuristic for multiple sequence alignment," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 499–514, Aug. 2016.

[31] M. Frutos, A. C. Olivera, and F. Tohmé, "A memetic algorithm based on a NSGAII scheme for the flexible job-shop scheduling problem," *Ann. Oper. Res.*, vol. 181, no. 1, pp. 745–765, 2010.

[32] Y.-F. Li, N. Pedroni, and E. Zio, "A memetic evolutionary multi-objective optimization method for environmental power unit commitment," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 2660–2669, Aug. 2013.

[33] F. Wang and Z. Zhu, "Global path planning of wheeled robots using a multi-objective memetic algorithm," in *Proc. Int. Conf. Intell. Data Eng. Autom. Learn. (IDEAL)*, 2013, pp. 437–444.

[34] H. Ishibuchi, T. Yoshida, and T. Murata, "Balance between genetic search and local search in memetic algorithms for multiobjective permutation flowshop scheduling," *IEEE Trans. Evol. Comput.*, vol. 7, no. 2, pp. 204–223, Apr. 2003.

[35] G. Ochoa, S. Verel, and M. Tomassini, "First-improvement vs. best-improvement local optima networks of NK landscapes," in *Proc. Int. Conf. Parallel Problem Solving Nat.*, 2010, pp. 104–113.

[36] J. Dubois-Lacoste, M. López-Ibá nez, and T. Stützle, "Pareto local search algorithms for anytime bi-objective optimization," in *Proc. Eur. Conf. Evol. Comput. Comb. Optim.*, 2012, pp. 206–217.

[37] M. M. Drugan and D. Thierens, "Stochastic Pareto local search: Pareto neighbourhood exploration and perturbation strategies," *J. Heuristics*, vol. 18, no. 5, pp. 727–766, 2012.

[38] L. Sheneman and J. A. Foster, "Estimating the destructiveness of crossover on binary tree representations," in *Proc. Annu. Conf. Companion Genet. Evol. Comput. (GECCO)*, vol. 2, 2006, pp. 1427–1428.

[39] S. Santander-Jiménez and M. Vega-Rodríguez, "A hybrid approach to parallelize a fast non-dominated sorting genetic algorithm for phylogenetic inference," *Concurrency Comput. Pract. Exp.*, vol. 27, no. 3, pp. 702–734, 2015.

[40] S. Santander-Jiménez and M. Vega-Rodríguez, "On the design of shared memory approaches to parallelize a multiobjective bee-inspired proposal for phylogenetic reconstruction," *J. Inf. Sci.*, vol. 324, no. 1, pp. 163–185, 2015.
[41] L. While, "A new analysis of the lebmeasure algorithm for calculating hypervolume," in *Proc. Int. Conf. Evol. Multi Criterion Optim.*, 2005, pp. 326–340.
[42] H. Matsuda, "Construction of phylogenetic trees from amino acid sequences using a genetic algorithm," in *Proc. Genome Informat. Workshop*, vol. 6, pp. 19–28, 1995.
[43] P. O. Lewis, "A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data," *Mol. Biol. Evol.*, vol. 15, no. 3, pp. 277–283, 1998.
[44] T. H. Reijmers, R. Wehrens, F. D. Daeyaert, P. J. Lewi, and L. M. C. Buydens, "Using genetic algorithms for the construction of phylogenetic trees: Application to G-protein coupled receptor sequences," *Biosystems*, vol. 49, no. 1, pp. 31–43, 1999.
[45] C. B. Congdon, "Gaphyl: An evolutionary algorithms approach for the study of natural evolution," in *Proc. Annu. Conf. Companion Genet. Evol. Comp.*, 2002, pp. 1057–1064.
[46] C. Cotta and P. Moscato, "Inferring phylogenetic trees using evolutionary algorithms," in *Proc. Int. Conf. Parallel Problem Solving Nat.*, 2002, pp. 720–729.
[47] S. Pirkwieser and G. Raidl, "Finding consensus trees by evolutionary, variable neighborhood search, and hybrid algorithms," in *Proc. Annu. Conf. Companion Genet. Evol. Comp.*, 2008, pp. 323–330.
[48] M. Villalobos-Cid, D. Vega-Araya, and M. Inostroza-Ponta, "Application of different multi-objective decision making techniques in the phylogenetic inference problem," in *Proc. Int. Conf. Chilean Comput. Sci. Soc. (SCCC)*, vol. 1, 2017, pp. 1–9.
[49] Z. Du, F. Lin, and U. W. Roshanb, "Reconstruction of large phylogenetic trees: A parallel approach," *Comput. Biol. Chem.*, vol. 29, no. 4, pp. 273–280, 2005.
[50] L. T. Nguyen, A. von Haeseler, H. A. Schmidt, and B. Q. Minh, "IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies," *Mol. Biol. Evol.*, vol. 32, no. 1, pp. 268–274, 2015.
[51] K. P. Schliep, "Phangorn: Phylogenetic analysis in R," *Bioinformatics*, vol. 27, no. 4, pp. 592–593, 2011.
[52] D. L. Swofford and W. P. Maddison, "Reconstructing ancestral character states under Wagner parsimony," *Math. Biosci.*, vol. 87, no. 2, pp. 199–229, 1987.
[53] N. D. Pattengale, E. J. Gottlieb, and B. M. Moret, "Efficiently computing the Robinson–Foulds metric," *Int. J. Comput. Biol.*, vol. 14, no. 6, pp. 724–735, 2007.
[54] J. E. Gallardo, C. Cotta, and A. J. Fernández, "Reconstructing phylogenies with memetic algorithms and branch-and-bound," in *Analysis of Biological Data: A Soft Computing Approach*. Singapore: World Sci., 2007, pp. 59–84.
[55] S. Bandyopadhyay, S. Saha, U. Maulik, and K. Deb, "A simulated annealing-based multiobjective optimization algorithm: AMOSA," *IEEE Trans. Evol. Comput.*, vol. 12, no. 3, pp. 269–283, Jun. 2008.
[56] S. Jiang, Y.-S. Ong, J. Zhang, and L. Feng, "Consistencies and contradictions of performance metrics in multiobjective optimization," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2391–2404, Dec. 2014.
[57] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Syst. Biol.*, vol. 52, no. 5, pp. 696–704, 2003.
[58] S. Kumar, G. Stecher, and K. Tamura, "MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets," *Mol. Biol. Evol.*, vol. 33, no. 7, pp. 1870–1874, 2016.
[59] H. Ishibuchi, Y. Nojima, and T. Doi, "Comparison between single-objective and multi-objective genetic algorithms: Performance comparison and performance measures," in *Proc. IEEE Int. Conf. Evol. Comput. (CEC)*, vol. 1, 2006, pp. 1143–1150.
[60] A. Rohatgi and ZlatanStanojevic. (2017). *WebPlotDigitizer: Version 3.11 of WebPlotDigitizer*. [Online]. Available: https://automeris.io/WebPlotDigitizer
[61] D. Drevon, S. R. Fursa, and A. L. Malcolm, "Intercoder reliability and validity of WebPlotDigitizer in extracting graphed data," *Behav. Modification*, vol. 41, no. 2, pp. 323–339, 2017.
[62] C. R. Carlini and R. Ligabue-Braun, "Ureases as multifunctional toxic proteins: A review," *J. Int. Soc. Toxicol.*, vol. 110, no. 1, pp. 90–109, 2016.
[63] R. Ligabue-Braun, F. C. Andreis, H. Verli, and C. R. Carlini, "3-to-1: Unraveling structural transitions in ureases," *Naturwissenschaften*, vol. 100, no. 5, pp. 459–467, 2013.
[64] J. Felsenstein. (2016). *PHYLIP (Phylogeny Inference Package) Version 3.6*. [Online]. Available: http://evolution.genetics.washington.edu/phylip.html
[65] G. Giribet, "Efficient tree searches with available algorithms," *Evol. Bioinformat.*, vol. 3, no. 3, pp. 341–356, 2007.
[66] B. Borguesan, M. Barbachan e Silva, B. Grisci, M. Inostroza-Ponta, and M. Dorn, "APL: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction," *Comput. Biol. Chem.*, vol. 59, pp. 142–157, Dec. 2015.

Manuel Villalobos-Cid received the B.Sc. degree in biomedical engineering from the Universidad de Valparaíso, Valparaíso, Chile, in 2008, and the Ph.D. degree in Ciencias de la Ingeniería Mención Informática from the Universidad de Santiago de Chile, Santiago, Chile, in 2017.

He has a professional profile related to the healthcare planning, control and production analysis of hospitals, researching with Hospital Barros Luco-Trudeau, San Miguel, Chile, and the Servicio de Salud Metropolitano Sur, Santiago, from 2012 to 2015. He is currently a Post-Doctoral Researcher with the Department of Informatics, Universidad de Santiago de Chile, and the Centre for Biotechnology and Bioengineering, Santiago. His current research interests include evolutionary computation, multiobjective optimization, data mining, and their applications to image processing, computational biology, and bioinformatics.

Márcio Dorn received the M.Sc. degree from the Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil, in 2008, and the Ph.D. degree in computer science from the Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, in 2012, both in computer science.

He is an Associate Professor with the Institute of Informatics, UFRGS, where he is also leading of the Structural Bioinformatics and Computational Biology Laboratory. In 2008, 2009, and 2017, he was a Research Associate with the Karlsruhe Institute of Technology, Karlsruhe, Germany. His current research interests include bioinformatics, structural bioinformatics, machine learning, metaheuristics, artificial Intelligence, and high-performance computing.

Prof. Dorn is a CNPq (Brazilian National Research Council) Advanced Fellow and an Alexander von Humboldt Research Fellow (AvH Germany) communities.

Rodrigo Ligabue-Braun received the Doctor of Science degree in cellular and molecular biology from the Federal University of Rio Grande do Sul, Porto Alegre, Brazil, in 2014.

He currently an Associate Professor with the Pharmaceutical Sciences Department, Universidade Federal de Ciências da Saúde de Porto Alegre, Porto Alegre. His research on venomous mammals has attracted readers both from the toxinology community and the general public alike. Aside from research, he has published chapters of textbooks in Brazil and also coordinated science outreach programs aiming low-income students from local communities. His current research interests include neglected toxins, moonlighting properties of proteins, and the effect of unstructured or metastable regions of polypeptides on the structure-function paradigm.

Mario Inostroza-Ponta received the Computer Engineering degree from the Universidad de Santiago de Chile, Santiago, Chile, in 2001, and the Ph.D. degree in computer science from the University of Newcastle, Callaghan, NSW, Australia, in 2008.

He is currently an Associate Professor with the Departamento de Ingeniería Informática, Universidad de Santiago de Chile. He spent a year as a Post-Doctoral Researcher with the Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil. His current research interests include the development of *ad-hoc* computational techniques to solve problems bioinformatics and other areas, using data mining, metaheuristics, and multiobjective optimization.