A novel approach based on multiobjective variable mesh optimization to Phylogenetics

Cristian Zambrano-Vega, Byron Oviedo Bayas, Stalin Carreño, Amilkar Puris, and Oscar Moncayo

***Abstract*—One of the most relevant problems in Bioinformatics and Computational Biology is the search and reconstruction of the most accurate phylogenetic tree that explains, as exactly as possible, the evolutionary relationships among species from a given dataset. Different criteria have been employed to evaluate the accuracy of evolutionary hypothesis in order to guide a search algorithm towards the best tree. However, these criteria may lead to distinct phylogenies, which are often conflicting among them. Therefore, a multi-objective approach can be useful. In this work, we present a phylogenetic adaptation of a multiobjective variable mesh optimization algorithm for inferring phylogenies, to tackle the phylogenetic inference problem according to two optimality criteria: maximum parsimony and maximum likelihood. The aim of this approach is to propose a complementary view of phylogenetics in order to generate a set of trade-off phylogenetic topologies that represent a consensus between both criteria. Experiments on four real nucleotide datasets show that our proposal can achieve promising results, under both multiobjective and biological approaches, with regard to other classical and recent multiobjective metaheuristics from the state-of-the-art.**

***Index Terms*—Multiobjective Optimization, Phylogenetic Inference, Evolutionary Computation, Bioinformatics.**

I. Introduction

The evolutionary history of mankind and all other living and extinct species on earth is a question which has been preoccupying mankind for centuries. Therefore, the construction of a “tree of life” comprising all living and extinct organisms on earth has been a fascinating and challenging idea since the emergence of evolutionary theory [1].

Typically, evolutionary relationships among organisms are represented by an evolutionary tree. Phylogenetic inference consists in finding the best tree that explains the genealogical relationships or evolutionary history of species from molecular sequences (DNA or protein data). The data used in this analysis usually come from aligned nucleotide or aminoacid sequences called Multiple Sequence Aligned [2], [3].

Various scientific fields can benefit thanks to the contributions of phylogenetic, such as evolutionary biology, physiology, ecology, paleontology, biomedicine, chemistry and others [4]. For all this, many scientists agree that phylogenetic inference is one of the most important research topics in Bioinformatics.

Handl et al. [5] discussed the applications of multiobjective optimization in several bioinformatics and computational biology problems, in this survey phylogenetic inference is one of the central problems in this area. Unfortunately, many interesting problems and algorithms in Bioinformatics, such as inference of perfect phylogenies or optimal multiple sequence alignment are NP-complete and computationally extremely intensive.

Recently several multiobjective proposed applied to phylogenetic inference have been published oriented to optimize trees under reconstruction criteria Maximum Parsimony and Maximum Likelihood: two bio-inspired techniques based in swarm intelligence algorithms: *MOABC* [4] an adaptation of the Artificial Bee Colony (ABC) and *Mo-FA* [6] a multiobjective adaptation of the novel Firefly Algorithm; and two other techniques based on the popular multi-objective metaheuristic the fast non-dominated sorting genetic algorithm (NSGAII): *PhyloMOEA* [7] and *MO-Phyl* [8] a hybrid OpenMP/MPI parallel technique.

In this work we present a phylogenetic adaptation of the multiobjective variable mesh optimization algorithm [9] called *PhyloMOVMO*, to infer phylogenetic trees optimizing two optimality criteria, simultaneously: the Maximum Parsimony and Maximum Likelihood, with the aim of allowing biologists to infer in a single run a set of trade-off phylogenetic topologies that represent a consensus between different points of both optimality criteria. In order to assess the performance of our proposal, we have carried out experiments on four nucleotide data sets extracted from the state-of-the-art, comparing the multiobjective and biological results with other popular and recient multiobjective metaheuritics applying multiobjective quality metrics. PhyloMOVMO has been implemented using funcionalities of the framework MO-Phylogenetics [10], a phylogenetic inference software tool with multi-objective evolutionary metaheuristics. The rest of the algorithms, the benchmark, the configurations and parameters files were taken from this framework.

The remainder of this paper is organized in the following way. In the Section II, we introduce concepts about the basis of phylogenetics, the complexity of the problem and the parsimony and likelihood methods. Section III explains the details about the PhyloMOVMO algorithm and the adaptation to phylogenetic inference. The followed experimental methodology to assess the performance of our proposal is described in the Section IV. The multiobjective and biological results are shown in Section V. And finally, Section VI summarizes some conclusions and future works about this topic.

Article history:
Received 13 September 2017
Accepted 28 November 2017

The authors are professors at Universidad Técnica Estatal de Quevedo, Quevedo, Los Ríos, Ecuador.
Email: czambrano, boviedo, sdcarreno, apuris, omoncayo}@uteq.edu.ec

II. Phylogenetic Inference fundamentals

Phylogenetic inference seeks to find the most accurate hypotheses about the evolution of species by combining statistical techniques and algorithmic procedures. In a phylogenetic analysis, we consider as input an alignment composed by n sequences of N characters (sites) that represent molecular characteristics of the organisms under review. Site values in the sequences belong to an alphabet Σ defined in accordance with the nature of the data, where for DNA sequences, Σ consists of four characters of the nucleotides {A, T, G, C} and for protein sequences, Σ consists of 20 characters of the amino acids {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. The output of the inference process is a tree-shaped structure $\tau = (V, E)$, where V represents the set of nodes in the tree τ and E the branches that connect related nodes V in the tree τ.

A. Complexity of the problem

The main computational problem of phylogenetic inference is the large number of possible topologies in the search space, which grows exponentially with the number of species to be analyzed.

Given n organisms, the number of possible binary unrooted trees is defined by equation 1 [11]:

$$|SS| = \prod_{i=1}^{n}(2_i - 5) = \frac{(2n-5)!}{2^{n-3}(n-3)!} \tag{1}$$

Due to the large number of possible combinations, the exhaustive methods become totally complex from a computational approach, when trying to infer phylogenies with more of ten species. Because of this “combinatorial explosion”, the phylogenetic inference is considered as NP-Hard problem, formally demonstrated both under an approach Maximum Parsimony [12] and Maximum Likelihood [13].

In the following subsections we will introduce the basis of two of the most used criteria-based methods for phylogenetic reconstruction: maximum parsimony and maximum likelihood analysis.

B. Maximum Parsimony Approach

Among the different hypotheses that explain the nature of a system, Occam’s reasoning suggests that the simplest hypothesis relative to a phenomenon must always be preferred. This statement is widely applied in a wide range of scientific domains, including Bioinformatics. The principle of parsimony is an analysis inspired by this reasoning.

The maximum parsimony method aims to find a tree that minimizes the number of character state changes (or evolutionary steps) that are needed to explain the data. It is preferred the tree whose topology implies a smaller amount of transformations at molecular level [14]. The problem of maximum parsimony is described as follows: Let D an input dataset containing n number of aligned sequences of species. Each aligned sequence has N sites (columns of characters), where d_{ij} is the state character of the sequence i at the site j. Given the τ tree with the set of nodes $V(\tau)$ and the set of branches $E(\tau)$, the parsimony value of the tree τ is defined as equation 2 [15].

$$PS(\tau) = \sum_{j=1}^{N} \sum_{(v,u)\in E(\tau)} w_j C(v_j, u_j) \tag{2}$$

where w_j refers to the weight of the site j, v_j and u_j are the character states of the nodes v and u in the site j for each branch (u, v) in τ, respectively, and C is the cost matrix, such that $C(v_j, u_j)$ is the cost to change the state v_j to state u_j.

In this work, we will use the algorithm proposed by Fitch [16] to compute the parsimony score of a phylogenetic tree.

Having defined the algorithm that minimizes $PS(\tau)$ for a tree τ, we have to find the tree τ^* such that $PS(\tau^*)$ is the score with the lowest value of parsimony in the whole space of trees.

C. Maximum Likelihood Approach

Likelihood is a statistical function that, applied to phylogenetics, indicates the probability that the evolutionary hypothesis involving a phylogenetic tree topology and a molecular evolution model Φ would give rise to the set of organisms observed in the input data D (set of aligned sequences) [15]. The maximum likelihood approach aims to find that tree representing the more likely evolutionary history of the organisms of the input data. It can be defined as follows: The likelihood of a phylogenetic tree, denoted by $L = P(D|\tau, \Phi)$, is the conditional probability of the data D given a tree τ and an evolutionary model Φ [14].

Given τ, $L = (\tau)$ can be defined as equation 3:

$$L(\tau) = \prod_{j=1}^{N} L_j(\tau) \tag{3}$$

where $L_j(\tau) = P(D_j|\tau, \Phi)$ is the likelihood in the site j, which is denoted as equation 4:

$$L_j(\tau) = \sum_{r_j} C_j(r_j, r).\pi_{r_j} \tag{4}$$

where r is the root node of τ, r_j refers to any possible state of r in the site j, π_{r_j} is the frequency of the state r_j and $C_j(r_j, r)$ is the conditional likelihood of the sub-tree rooted by r. Specifically, $C_j(r_j, r)$ is the probability of everything that is observed from the root node r to the leaves of the tree τ, in the site j and given r, has state r_j. Let u and v the descendant nodes next to r, $C_j(r_j, r)$ can be formulated as equation 5:

$$C_j(r_j, r) = \left[\sum_{u_j} C_j(u_j, u).P(r_j, u_j, t_{ru})\right]\left[\sum_{v_j} C_j(v_j, v).P(r_j, v_j, t_{rv})\right] \tag{5}$$

where u_j y v_j refers to any state of the nodes u y v, respectively. t_{ru} and t_{rv} are the branch lengths that join the node r with the nodes v and u, respectively. $P(r_j, u_j, t_{ru})$ is the probability of change from the state r_j to the state u_j while the evolutionary time t_{ru}. Similarly, $P(r_j, v_j, t_{rv})$ is the probability of change from the state r_j to the state v_j in the time t_{rv}. Both probabilities are provided by the evolutionary model Φ.

In this work, to calculate L we will use the method proposed by Felsenstein [14], where L is obtained by a post-order traversal in τ. Usually, it is convenient to use logarithmic values of L, so that the equation (3) can be redefined as equation 6:

$$\ln L(\tau) = \prod_{j=1}^{N} \ln L_j(\tau) \tag{6}$$

III. A MULTIOBJECTIVE VARIABLE MESH OPTIMIZATION APPROACH FOR PHYLOGENETIC INFERENCE

In this section we describe the main features of our proposal, a phylogenetic adaptation of the Multiobjective variable mesh optimization algorithm (MOVMO) proposed by [9]. Algorithm 1 shows the PhyloMOVMO's general workflow. The parameters: Mesh size P, Neighborhood size k, Number maximun of evaluations C, Maximun archive size S are the same of the MOVMO algorithm. We have included the input dataset to infer phylogenies: the multiple sequence alignments with the set of aligned sequences, the initial phylogenetic trees, and the evolutionary model parameters for each dataset, which can be computed by using jModelTest [17]. The representation of the individuals is based on the standard tree template codification. The crossover operator is the Prune-Delete-Graft (PDG) recombination method [18]. The output of the algorithm will be a set of non-dominated solutions L (Pareto set approximation) that describes trade-off phylogenetic topologies.

The algorithm starts by generating the initial mesh Pop_0 and initializing the leaders archive L (using Algorithm 2) with all the non-dominated solutions in Pop_0 (Lines 1 and 2). These initial solutions are assigned randomly from a repository composed by phylogenies generated by a bootstrap analysis [14]. For each node n_i of the current mesh Pop, the following steps are carried out:

1) The best node n_i^* among the n_i's k nearest neighbors in the decision variable space is selected. The distance between the nodes (phylogenetic trees) is calculated according to the Robinson-Foulds metric and the best node is selected according to the multiobjective dominance criterion (Line 5).
2) If the local optimum dominates n_i, a new node n_l is generated by applying TreeCrossover operator using n_i^* and n_i (Line 7); otherwise n_i is the local optimum itself (Line 8).
3) A global leader n_g from the archive L is selected through Binary Tournament selection operator (Line 10). Two non-dominated solutions from L are randomly picked and the one with largest crowding distance in L is selected.
4) A TreeCrossover operator is applied over the global leader n_g with the local optimum n_l (Line 11) to generate a new solution n_x, which contains subtrees of both topologies (mesh nodes).

Algorithm 1: Phylogenetic Multiobjective Variable Mesh Optimization (PhyloMOVMO)

```
Input: mesh size $P$, neighborhood size $k$, number max.
       of evaluations $C$, maximun archive size $S$
Data: multiple sequence alignment, initial trees and
      evolutionary model
Result: An approximation $L$ of the true Pareto set $L^*$
1  $Pop \leftarrow$ Initialize_Evaluate_Population($P$);
2  $L$=Initialize archive with each mesh node $n_i$ by
     Algorithm 2;
3  $c \leftarrow 1$;
4  while node $n_i$ in the current mesh $Pop$ do
5      $n_i^* \leftarrow$ the best among the $k$ neighbors of $n_i$
6      if $n_i^* \prec n_i$ then
7          $n_l \leftarrow$ PDGTreeCrossover($n_i^*$,$n_i$);
8      else
9          $n_l \leftarrow n_i$
10     $n_g \leftarrow$ Select a global leader from $L$
         (BinaryTournament_Selection);
11     $n_x \leftarrow$ PDGTreeCrossover($n_l$,$n_g$);
12     $n_x$:PhylogeneticOptimization(PPN&PLL);
13     evaluateFitness($n_x$);
14     add $n_x$ to the Pareto set approximation $L$ (see
         Algorithm 2);
15     if $n_x \preceq n_i$ then
16         Replace $n_i$ with $n_x$ in the current population $Pop$
17     $c \leftarrow c + 1$;
18 return $L$
```

5) A phylogenetic optimization method is applied on n_x (Line 12), a Local Search provided by MO-Phylogenetics [10] based on two highly optimized techniques to explore the tree space, pllRearrangeSearch [19] and PPN [20], to optimize the likelihood and parsimony objectives, respectively.
6) The new n_x node is evaluated and, if is a new non-dominated solution, is added to the Pareto set approximation L. All dominated solutions by n_x are deleted in L (Line 13 and 14).
7) Finally, if n_i is dominated by n_x, it is replaced with n_x in the current mesh Pop (Line 16).

PhyloMOVMO returns the leaders archive L as the approximation of the Pareto optimal set found.

Algorithm 2 describes the addition of a new mesh node n_x to the bounded leader archive L. First, all nodes in L that are dominated by the incoming solution are deleted from the archive prior to n_x's addition. If the archive reached its maximum size, we drop the node with the lowest crowding distance. This ensures that a well-spread set of non-dominated solutions is maintained in L.

Evolutionary Crossover Operator

A wide range of recombination operators can be found in the literature [21], [22]. We have used in our proposal the Prune-Delete-Graft (PDG) recombination operator [18] available in

Algorithm 2: Add the n_x solution to the leader archive L)

```
Input: Solution $n_x$, archive $L$
Result: Archive $L$
foreach $n_j$ of $L$ do
    if $n_x \prec n_j$ then
        $L \leftarrow L - n_j$ ;    /* remove $n_j$ from the archive */
    else if $n_x = n_j \parallel n_j \preceq n_x$ then
        exit ;                      /* discard $n_x$ */
$L \leftarrow L \cup n_x$ ;    /* add $n_x$ to the archive */
if $L : size() \succ L : maxSize()$ then
    recompute crowding distances in $L$;
    $L \leftarrow L -$ {L:worstByCrowdingDistance} ;
        /* remove most crowded solution */
```

MO-Phylogenetics. This operator takes a random subtree from one of the tree and inserts it in the other tree at a randomly selected insertion point, deleting duplicated species. Fig. 1 ilustrates the operator.

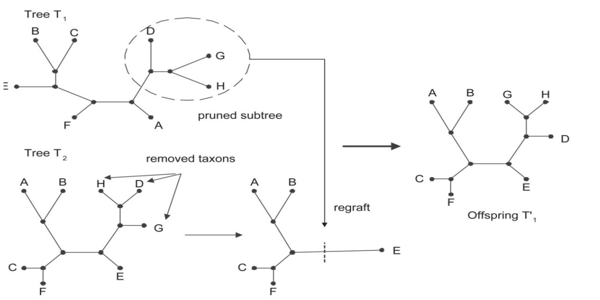


Fig. 1. Example of the Prune-Delete-Graft crossover operator.

IV. Experimental Methodology

In this section, we summarize the experimental methodology used to assess the performance achieved by our proposal PhyloMOVMO.

To comparate the results of our proposal, we have selected three representative multiobjective algorithms of the state-of-the-art, the classical reference NSGA-II and two other moderm techniques MOEA/D and SMS-EMOA, which are representative techniques of decomposition and indicator-based algorithms, respectively.

- NSGA-II [23] is a generational genetic algorithm based on generating new individuals from the original population by applying the typical genetic operators (selection, crossover and mutation). A ranking procedure is applied to promote convergence, while a density estimator (the crowding distance) is used to enhance the diversity of the set of found solutions.
- MOEA/D [24] is based on decomposing a multi-objective optimization problem into a number of scalar optimization subproblems, which are optimized simultaneously, only using information from their neighboring subproblems. This algorithm also applies a mutation operator to the solutions.
- SMS-EMOA [25] is a steady-state evolutionary algorithm that uses a selection operator based on the hyper-volume measure combined with the concept of non-dominated sorting.

We have used four multiobjective quality indicators: the Hypervolume (I_{HV}) and the Inverted Generational Distance Plus or IGD$^+$ (I_{IGD^+}) to take into account both the convergence and diversity of the Pareto front approximations, the Unary Additive Epsilon ($I_{\epsilon+}$) and the Spread or Δ (I_Δ) indicators, that are used as a complement to measure the degree of convergence and diversity, respectively. As we are dealing with real-world optimization problems, the Pareto fronts to calculate these two metrics are not known, so we have generated a reference Pareto front for each nucleotide dataset by combining all the non-dominated solutions computed in all the executions of all the algorithms. This strategy allows to make a relative performance assessment of the metaheuristics, because if the behavior of all the compared techniques is poor we know which of them yields the best fronts, but we do not know if they are near or far from the true Pareto front.

The experiments were carried out on four nucleotide data sets from the literature [26]: *rbcL_55* 55 sequences with 1314 nucleotides per sequence of the rbcL gene, *mtDNA_186* 186 sequences with 16608 nucleotides per sequence of human Mt DNA, *RDPII_218* 218 sequences with 4182 nucleotides per sequence of prokaryotic RNA and *ZILLA_500* 500 sequences with 759 nucleotides per sequence of rbcL plastid gene, under the reliable General Time Reversible (GTR+Γ) evolutionary model [27]. For each combination of algorithm and nucleotide dataset problem we have carried out 20 independent runs, and we report the median, $\tilde{x}$, and the interquartile range, IQR, as measures of location (or central tendency) and statistical dispersion, respectively, for every considered indicators. When presenting the obtained values in tables, we emphasize with a dark gray background the best result for each problem, and a clear grey background is used to indicate the second best result; this way, we can see at a glance the most salient algorithms. To check if differences in the results are statistically significant, we have applied the unpaired Wilcoxon rank-sum test. A confidence level of 95% (i.e., significance level of 5% or *p*-value under 0.05) has been used in all cases. The results of these tests have been summarized in tables where each cell contains the results of this test for a pair of algorithms. Three different symbols are used: "–" indicates that there is no statistical significance between these algorithms, "▲" means that the algorithm in the row has yielded better results than the algorithm in the column with statistical confidence, and "▽" is used when the algorithm in the column is statistically better than the algorithm in the row.

All the algorithms use the same parameters, the crossover and mutation probabilities are 0.8 and 0.2. The population size is 100. The initial population is generated by using a set of user phylogenetic trees performed by bootstrap analysis [14]. The parameters of the evolutionary model are computed by jModelTest [17].

TABLE I
MEDIAN AND INTERQUARTILE RANGE IQR OF THE VALUES OF THE $I_{\epsilon+}$ INDICATOR.

	PhyloMOVMO	NSGAII	MOEAD	SMSEMOA
rbcL_55	$2.10e-01_{1.4e-01}$	$1.01e+00_{5.0e-01}$	$1.75e-01_{4.1e-02}$	$2.50e-01_{8.5e-02}$
mtDNA_186	$3.33e-01_{1.5e-01}$	$3.38e-01_{1.1e-01}$	$3.89e-01_{1.7e-01}$	$4.44e-01_{2.1e-01}$
RDPII_218	$1.18e-01_{3.6e-02}$	$1.36e-01_{2.7e-02}$	$1.59e-01_{5.6e-02}$	$1.51e-01_{4.7e-02}$
ZILLA_500	$7.68e-01_{3.8e-01}$	$9.33e-01_{2.0e-01}$	$8.13e-01_{3.7e-01}$	$9.38e-01_{3.0e-01}$

TABLE II
MEDIAN AND INTERQUARTILE RANGE IQR OF THE VALUES OF THE I_{Δ} INDICATOR.

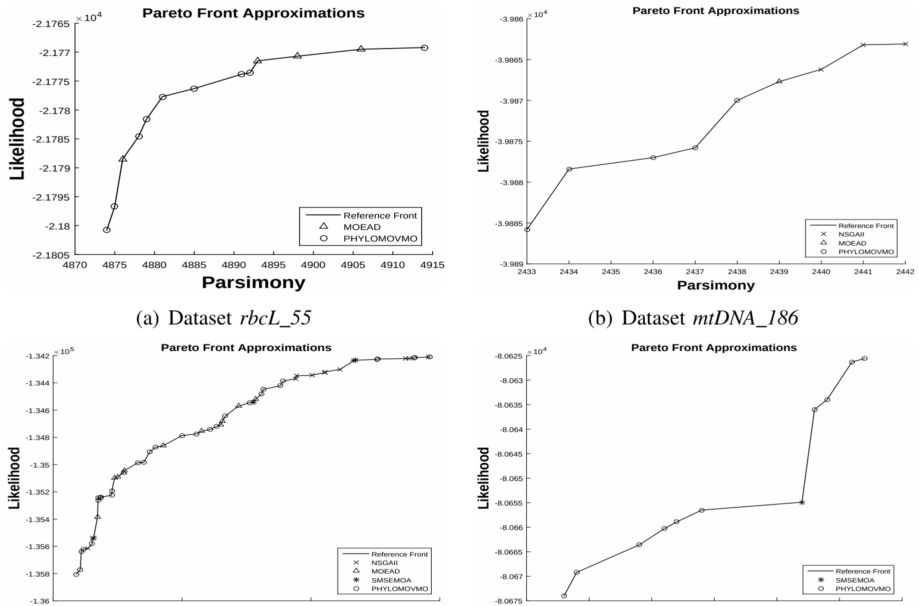
	PhyloMOVMO	NSGAII	MOEAD	SMSEMOA
rbcL_55	$9.91e-01_{2.5e-01}$	$1.01e+00_{3.4e-01}$	$1.14e+00_{2.8e-01}$	$8.76e-01_{3.8e-01}$
mtDNA_186	$1.31e+00_{3.9e-01}$	$7.97e-01_{5.0e-01}$	$1.30e+00_{1.5e-01}$	$1.13e+00_{7.7e-01}$
RDPII_218	$8.89e-01_{1.8e-01}$	$7.99e-01_{6.8e-02}$	$1.13e+00_{9.4e-02}$	$9.11e-01_{1.8e-01}$
ZILLA_500	$1.09e+00_{1.6e-01}$	$8.44e-01_{1.1e-01}$	$1.16e+00_{8.4e-02}$	$9.74e-01_{2.4e-01}$

TABLE III
MEDIAN AND INTERQUARTILE RANGE IQR OF THE VALUES OF THE I_{HV} INDICATOR.

	PhyloMOVMO	NSGAII	MOEAD	SMSEMOA
rbcL_55	$6.34e-01_{1.7e-01}$	$0.00e+00_{0.0e+00}$	$6.83e-01_{5.5e-02}$	$5.81e-01_{1.3e-01}$
mtDNA_186	$3.07e-01_{1.3e-01}$	$2.56e-01_{1.1e-01}$	$2.75e-01_{1.6e-01}$	$2.40e-01_{1.6e-01}$
RDPII_218	$6.18e-01_{4.2e-02}$	$6.08e-01_{5.5e-02}$	$5.87e-01_{8.9e-02}$	$5.99e-01_{4.1e-02}$
ZILLA_500	$1.57e-02_{1.0e-01}$	$0.00e+00_{1.1e-02}$	$2.88e-03_{8.7e-02}$	$0.00e+00_{3.1e-02}$

TABLE IV
MEDIAN AND INTERQUARTILE RANGE IQR OF THE VALUES OF THE I_{IGD+} INDICATOR.

	PhyloMOVMO	NSGAII	MOEAD	SMSEMOA
rbcL_55	$1.10e-01_{1.2e-01}$	$9.35e-01_{5.2e-01}$	$8.71e-02_{4.0e-02}$	$1.48e-01_{7.7e-02}$
mtDNA_186	$1.90e-01_{1.2e-01}$	$2.33e-01_{8.9e-02}$	$2.26e-01_{2.0e-01}$	$2.37e-01_{2.2e-01}$
RDPII_218	$6.87e-02_{2.5e-02}$	$7.55e-02_{3.5e-02}$	$8.57e-02_{5.0e-02}$	$7.77e-02_{3.2e-02}$
ZILLA_500	$5.56e-01_{6.3e-01}$	$8.25e-01_{3.2e-01}$	$6.97e-01_{4.7e-01}$	$5.99e-01_{2.0e-01}$



(a) Dataset *rbcL_55*

(b) Dataset *mtDNA_186*

(c) Dataset *RDPII_218*

(d) Dataset *ZILLA_500*

Fig. 2. Reference Pareto fronts and best Pareto front approximations obtained by all the algorithms (PhyloMOVMO, NSGAII, MOEA/D and SMSEMOA) over 20 independent runs solving the nucleotide datasets a) *rbcL_55*, b) *mtDNA_186*, c) *RDPII_218* and d) *ZILLA_500*.

V. Empirical Results and Statistical Analysis

In this section we analyze the PhyloMOVMO's multiobjective and biological performance compared to NSGA-II, MOEA/D and SMSEMOA solving four nucleotide datasets from benchmark based on the experimental methodology described in Section IV.

Multiobjective results

The median values, $\tilde{x}$, and the interquartile range, IQR of the quality indicators $I_{\epsilon+}$, $\mathrm{I}_{\triangle}$, I_{HV} and I_{IGD+} are reported in the Tables I, II, III and IV, respectively. We have to consider the highest values of I_{HV} and I_{IGD+} and the lowest of $I_{\epsilon+}$ and $\mathrm{I}_{\triangle}$

The results of $I_{\epsilon+}$, I_{HV} and I_{IGD+} show that PhyloMOVMO obtains the best median values for all the datasets, except for the *rcbL_55* instance, where MOEA/D shows a better performance. And for the results of $\mathrm{I}_{\triangle}$ occurs the same, NSGAII obtains the best median values for all the datasets, except for the *rcbL_55* instance, where SMSEMOA shows a better performance.

Pareto Front approximations

To ilustrate graphically the multiobjective quality indicators results, we ilustrate in the Figure 2 the reference Pareto front (described in Section IV), and the best Pareto front approximations obtained by all the algorithms (PhyloMOVMO, NSGAII, MOEA/D and SMSEMOA) over 20 independent runs solving the nucleotide datasets *rbcL_55*, *mtDNA_186*, *RDPII_218* and *ZILLA_500*.

We can observe that all reference Pareto fronts of the Figure 2, are mostly conformed by the non-dominated solutions (phylogenies) of the Pareto front approximations of PhyloMOVMO, considering only a few solutions of MOEA/D and SMSEMOA for the *rbcL_55* and *ZILLA_500* datasets,respectively. Furthemore, in the Figure 2c we can observe a high competitive performance that exists between all the algorithms solving the *RDPII_218* nucleotide dataset.

Furthermore, the Figure 3 shows the reference Front and the Pareto front approximations of each algorithm of each nucleotide dataset, from the best values of I_{HV} and I_{IGD+} indicators, respectively, considering that both take into account the convergence and diversity of the Pareto front approximations. All these Pareto fronts approximations confirm the multiobjective quality indicators results of the Tables I, II, III and IV.

Statistical Analysis

The Tables V, VI, VII and VIII show the the Wilcoxon rank-sum test results. These results confirm that PhyloMOVMO yielded better performance at 95% significance level on the $I_{\epsilon+}$, I_{HV} and I_{IGD+} values for the datasets *mtDNA_186*, *RDPII_218* and *ZILLA_500*, except for the *rcbL_55* instance where MOEA/D reports a better performance overall the algorithms. Furthermore, these results confirm the best performance of NSGAII for the $\mathrm{I}_{\triangle}$ values, and although the SMSEMOA reports the best performance over the *rcbL_55* instance in this indicator, the Wilcoxon test indicates that they are not statistically significant with the rest of the algorithms except for MOEA/D.

TABLE V
Results of the Wilcoxon rank-sum test for the $I_{\epsilon+}$ values for the datasets *rbcL_55*, *mtDNA_186*, *RDPII_218* and *ZILLA_500*.

	NSGAII	MOEAD	SMSEMOA
PhyloMOVMO	▲ – – ▲	▽ ▲ ▲ –	▲ ▲ ▲ ▲
NSGAII		▽ ▲ ▲ ▽	▽ ▲ ▲ –
MOEAD			▲ – – ▲

TABLE VI
Results of the Wilcoxon rank-sum test for the $\mathrm{I}_{\triangle}$ values for the datasets *rbcL_55*, *mtDNA_186*, *RDPII_218* and *ZILLA_500*.

	NSGAII	MOEAD	SMSEMOA
PhyloMOVMO	– ▽ ▽ ▽	– – ▲ ▲	– – – –
NSGAII		– ▲ ▲ ▲	– – – ▲
MOEAD			▽ – ▽ ▽

TABLE VII
Results of the Wilcoxon rank-sum test for the I_{HV} values for the datasets *rbcL_55*, *mtDNA_186*, *RDPII_218* and *ZILLA_500*.

	NSGAII	MOEAD	SMSEMOA
PhyloMOVMO	▲ ▲ – ▲	▽ – ▲ ▲	▲ ▲ ▲ ▲
NSGAII		▽ – ▲ ▽	▽ – ▲ –
MOEAD			▲ – – ▲

TABLE VIII
Results of the Wilcoxon rank-sum test for the I_{IGD+} values for the datasets *rbcL_55*, *mtDNA_186*, *RDPII_218* and *ZILLA_500*.

	NSGAII	MOEAD	SMSEMOA
PhyloMOVMO	▲ ▲ – ▲	– ▲ ▲ ▲	▲ ▲ ▲ –
NSGAII		▽ ▽ ▲ –	▽ – ▲ –
MOEAD			▲ ▲ – –

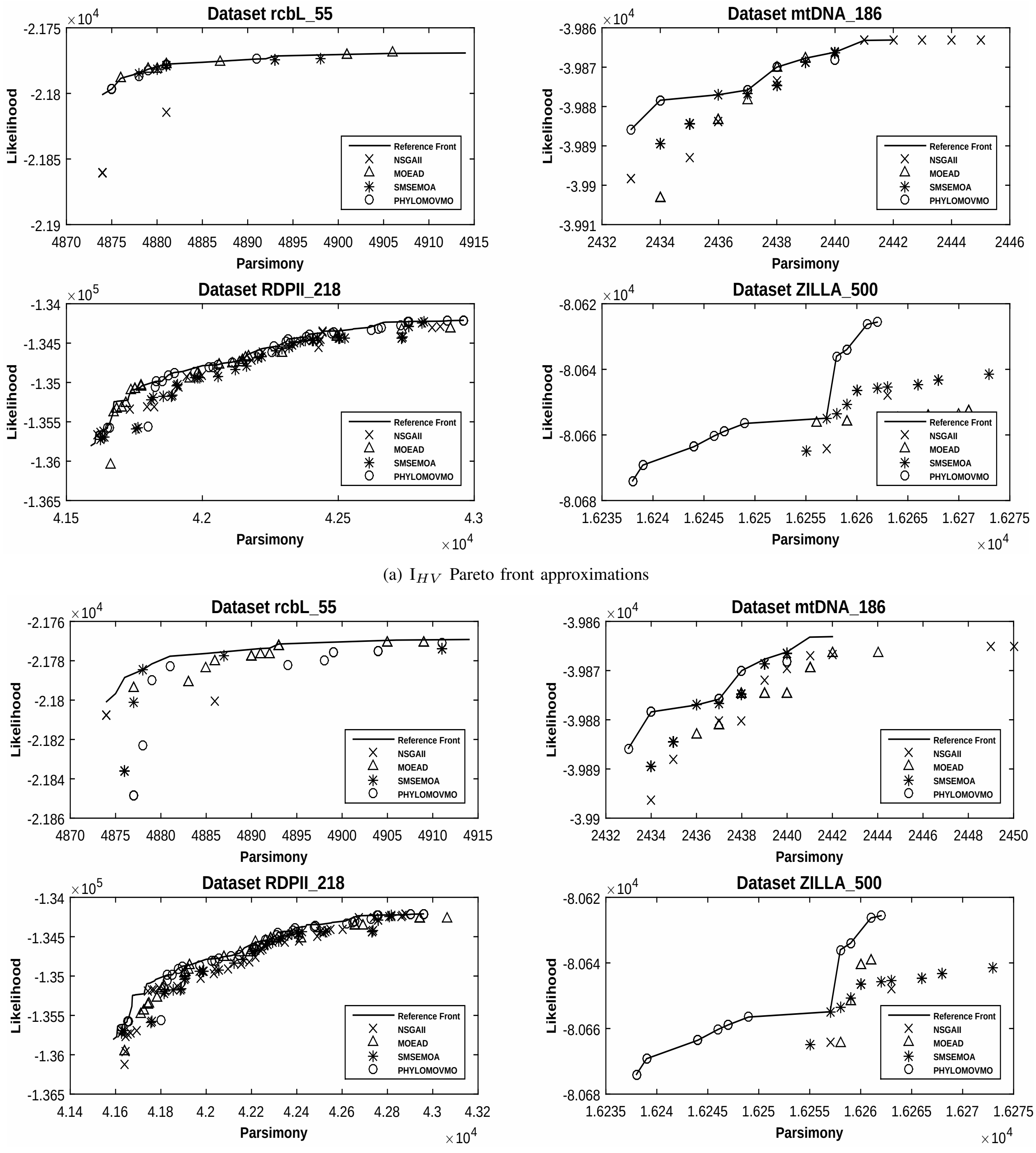
Biological results

The Table IX shows the best maximum parsimony and maximum likelihood scores obtanied by all the algorithms solving the four nucleotide datasets.

TABLE IX
Phylogenetic results. Comparing Parsimony and Likelihood scores of PhyloMOVMO with other multiobjective metaheuristics.

Dataset	**PhyloMOVMO**		**NSGAII**		**MOEAD**		**SMSEMOA**	
	Par.	**Lik.**	**Par.**	**Lik.**	**Par.**	**Lik.**	**Par.**	**Lik.**
rbcL_55	**4874**	**-21769.22**	4874	-21800.81	4874	-21769.53	4874	-21773.63
mtDNA_186	**2133**	-39865.52	2433	**-39863.11**	2434	-39866.60	2434	-39864.73
RDPII_218	**41589**	**-134210.40**	41613	-134211.03	41634	-134238.30	41618	-134224.12
ZILLA_500	**16238**	**-80625.60**	16251	-80630.91	16251	-80639.42	16250	-80628.03

We can observe that our proposal achieves a significant improvement with regard to the parsimony and likelihood scores reported by the other algorithms, except for the *rbcL_55* dataset where all the algorithms generate the same parsimony scores and for the *mtDNA_186* dataset where NSGAII performs a better likelihood score overall the algorithms.



(a) I_{HV} Pareto front approximations

(b) I_{IGD+} Pareto front approximations

Fig. 3. Pareto front approximations from the best I_{HV} and I_{IGD+} values obtained by all the algorithms (PhyloMOVMO, NSGAII, MOEA/D and SMSEMOA) over 20 independent runs resolving each nucleotide dataset.

Run-time analysis

Table X shows the computational processing times (in seconds) required to perform a complete execution of each algorithm (PhyloMOVMO, NSGAII, MOAED and SMSEMOA) using a single thread, making a phylogenetic analysis on each considered nucleotide dataset (rbcL_55, mtDNA_186, RDPII_218 and ZILLA_500).

TABLE X
SEQUENTIAL PROCESSING TIMES (SECS).

Dataset	PhyloMOVMO	NSGAII	MOAED	SMSEMOA
rbcL_55	5230.02	6376.34	22039.08	7500.19
mtDNA_186	39987.29	41870.29	47730.54	32362.21
RDPII_218	82901.13	79871.19	96085.18	84399.27
ZILLA_500	84090.31	82981.27	99098.10	86780.32

We can observe that the run-time of the algorithms is very expensive, specially for the large-scale datasets, requiring hours for mtDNA_186 dataset and almost a whole day for the RDPII_218 and ZILLA_500 datasets. The single-thread version of PhyloMOVMO and NSGAII perform a faster execution than the other algorithms.

VI. CONCLUSIONS AND FUTURE WORKS

In this paper, we have presented the PhyloMOVMO algorithm, a novel approach based on a multiobjective variable mesh optimization technique for inferring phylogenies, optimizing both parsimony and likelihood criteria simultaneously. Unlike to other multiobjective proposals applied to phylogenetic inference, the solutions selection based on the Robinson-Foulds distance metric, adds a new perspective to the exploration of the tree-space.

With the aim of evaluating its multiobjective and biological performance, we have carried out experiments on four nucleotide datasets and applying multiobjective quality indicators with other classical and recent multiobjective metaheuristics. With the purpose of making a fair comparison, all the algorithms were configured using the same parameters.

The obtained results reveal that in the context of the adopted parameter settings, the experimentation methodology, and the solved datasets, PhyloMOVMO shows a very competitive performance, under both multiobjective and biological approaches. The reference Pareto fronts of each dataset, are almost totally composed by the non-dominated solutions generated by PhyloMOVMO. Furthermore, the values of the multiobjetive quality indicators shows a promising perfomance of our proposal and to confirm these results, a Wilcoxon rank-sum analysis indicates the significant statistically differences of our proposal. Finally, under biological approach, PhyloMOVMO obtanied the best parsimony and likelihood scores overall data sets, except for the mtDNA_186 dataset where NSGAII provided a better likelihood score.

In summary, preliminary results have shown that PhyloMOVMO can make relevant contributions to phylogenetic inference. Moreover, there are several aspects that can be investigated to improve the current approach, such as: a parameter sensitivity study (including the use of different phylogenetic optimization methods), improve the funcionality of the recombination operator, add new evolutionary models to support protein data sets and, PhyloMOVMO requires several hours to find acceptable Pareto-solutions if initial trees are poorly estimated, so performance can be improved using the benefits of shared-memory and distributed-memory programming paradigms to efficiently inferring phylognies of large-size sequences data sets with a lot of number of species.

REFERENCES

[1] A. Stamatakis, "Distributed and Parallel Algorithms and Systems for Inference of Huge Phylogenetic Trees based on the Distributed and Parallel Algorithms and Systems for Inference of Huge Phylogenetic Trees based on the Maximum Likelihood Method," Ph.D. dissertation, Technische Universität München, Germany, 10 2004.

[2] C. Zambrano-Vega, A. J. Nebro, J. J. Durillo, J. García-Nieto, and J. Aldana-Montes, "Multiple sequence alignment with multiobjective metaheuristics. a comparative study," *International Journal of Intelligent Systems*, vol. 32, no. 8, pp. 843–861, 2017. [Online]. Available: http://dx.doi.org/10.1002/int.21892

[3] C. Zambrano-Vega, A. J. Nebro, J. García-Nieto, and J. Aldana-Montes, "Comparing multi-objective metaheuristics for solving a three-objective formulation of multiple sequence alignment," *Progress in Artificial Intelligence*, pp. 1–16, 2017. [Online]. Available: http://dx.doi.org/10.1007/s13748-017-0116-6

[4] 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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0303264713001615

[5] J. Handl, D. B. Kell, and J. Knowles, "Multiobjective optimization in bioinformatics and computational biology." *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 4, no. 2, pp. 279–92, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17473320

[6] S. Santander-Jiménez and M. A. Vega-Rodríguez, "A multiobjective proposal based on the firefly algorithm for inferring phylogenies," in *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Springer, 2013, pp. 141–152.

[7] W. Cancino and A. C. Delbem, "A Multi-objective Evolutionary Approach for Phylogenetic Inference," in *Evolutionary Multi-Criterion Optimization*, ser. Lecture Notes in Computer Science, S. Obayashi, K. Deb, C. Poloni, T. Hiroyasu, and T. Murata, Eds. Springer Berlin Heidelberg, 2007, vol. 4403, pp. 428–442. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-70928-2_34

[8] S. Santander-Jiménez and M. A. Vega-Rodríguez, "A hybrid approach to parallelize a fast non-dominated sorting genetic algorithm for phylogenetic inference," *Concurrency and Computation: Practice and Experience*, vol. 27, no. 3, pp. 702–734, 2014. [Online]. Available: http://doi.wiley.com/10.1002/cpe.3269

[9] Y. Salgueiro, J. L. Toro, R. Bello, and R. Falcon, "Multiobjective variable mesh optimization," *Annals of Operations Research*, pp. 1–25, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10479-016-2221-5

[10] C. Zambrano-Vega, A. J. Nebro, and J. Aldana-Montes, "Mo-phylogenetics: a phylogenetic inference software tool with multiobjective evolutionary metaheuristics," *Methods in Ecology and Evolution*, vol. 7, no. 7, pp. 800–805, 2016. [Online]. Available: http://dx.doi.org/10.1111/2041-210X.12529

[11] A. Edwards, L. Cavalli-Sforza, V. Heywood *et al.*, "Phenetic and phylogenetic classification," *Systematic Association Publication No. 6*, pp. 67–76, 1964.

[12] W. H. Day, D. S. Johnson, and D. Sankoff, "The computational complexity of inferring rooted phylogenies by parsimony," *Mathematical biosciences*, vol. 81, no. 1, pp. 33–42, 1986.

[13] B. Chor and T. Tuller, "Maximum likelihood of evolutionary trees: hardness and approximation," *Bioinformatics*, vol. 21, no. suppl 1, pp. i97–i106, 2005.

[14] J. Felsenstein, *Inferring Phylogenies*. Palgrave Macmillan, 2004. [Online]. Available: http://books.google.fr/books?id=GI6PQgAACAAJ

[15] D. Swofford, G. Olsen, P. Waddell, and D. Hillis, "Phylogeny reconstruction," in *Molecular Systematics*, 3rd ed. Sinauer, 1996, ch. 11, pp. 407–514.

[16] W. M. Fitch, "Toward defining the course of evolution: Minimum change for a specific tree topology," *Systematic Biology*, vol. 20, no. 4, pp. 406–416, 1971. [Online]. Available: http://sysbio.oxfordjournals.org/content/20/4/406.abstract

[17] D. Darriba, G. L. Taboada, R. Doallo, and D. Posada, "jmodeltest 2: more models, new heuristics and parallel computing," *Nature methods*, vol. 9, no. 8, pp. 772–772, 2012.

[18] C. Cotta and P. Moscato, "Inferring phylogenetic trees using evolutionary algorithms," in *International Conference on Parallel Problem Solving from Nature*. Springer, 2002, pp. 720–729.

[19] T. Flouri, F. Izquierdo-Carrasco, D. Darriba, A. Aberer, L.-T. Nguyen, B. Minh, A. Von Haeseler, and A. Stamatakis, "The Phylogenetic Likelihood Library," *Systematic Biology*, vol. 64, no. 2, pp. 356–362, 2015.

[20] A. Goëffon, J.-M. Richer, and J.-K. Hao, "Progressive tree neighborhood applied to the maximum parsimony problem." *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 5, no. 1, pp. 136–45, 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18245882

[21] H. Matsuda, "Construction of Phylogenetic Trees from Amino Acid Sequences using a Genetic Algorithm," *Sciences, Computer*, p. 560, 1995.

[22] C. B. Congdon, "Gaphyl: An Evolutionary Algorithms Approach For The Study Of Natural Evolution," in *GECCO 2002: Proceedings of the Genetic and Evolutionary Computation Conference*, 2002, pp. 1057–1064.

[23] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[24] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Trans. Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, 2007.

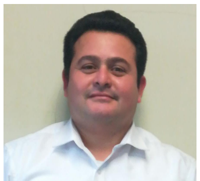
[25] N. Beume, B. Naujoks, and M. Emmerich, "Sms-emoa: Multiobjective selection based on dominated hypervolume," *European Journal of Operational Research*, vol. 181, no. 3, pp. 1653–1669, 2007.

[26] W. Cancino and A. Delbem, "A multi-criterion evolutionary approach applied to phylogenetic reconstruction," in *New Achievements in Evolutionary Computation*, P. Korosec, Ed. Rijeka: InTech, 2010, ch. 06. [Online]. Available: http://dx.doi.org/10.5772/8051

[27] C. Lanave, G. Preparata, C. Sacone, and G. Serio, "A new method for calculating evolutionary substitution rates," *Journal of molecular evolution*, vol. 20, no. 1, pp. 86–93, 1984.

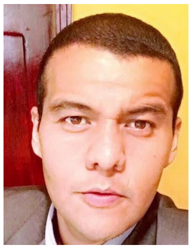
Cristian Zambrano-Vega Docente investigador de la Carrera de Ingeniería en Sistemas de la Facultad de Ciencias de la Ingienería de la Universidad Técnica Estatal de Quevedo. Doctor en Ingeniería del Software e Inteligencia Artificial de la Universidad de Málaga. Su línea de investigación abarca las técnicas de optimización multiobjetivo aplicadas a la Inferencia Filogenética y al Alineamiento Múltiple de Secuencias. Email czambrano@uteq.edu.ec

Byron Oviedo Bayas Director del Departamento de Investigación y Docente titular principal de la Universidad Técnica Estatal de Quevedo. Doctor en el programa oficial de Tecnologías de la Información y Comunicación de la Universidad de Granada - España. Email: boviedo@uteq.edu.ec.



Stalin Carreño Sandoya Ingeniero en Sistemas y Master en Conectividad y Redes de Ordenadores. Docente de la Unidad de Estudios a Distancia. Líder de la Unidad de Tecnologías de la Información y Comunicación de la Universidad Técnica Estatal de Quevedo. Email: sdcarreno@uteq.edu.ec

Amilkar Puris Cáceres Ph.D. en Ciencias Técnicas por la Universidad Marta Abreu de las Villas, Cuba. Sus principales investigaciones han sido en el área de las Metaheurísticas Poblacionales para la solución de problemas complejos. Actualmente se desempeña como docente e investigador en la Universidad Técnica Estatal de Quevedo. Email: apuris@uteq.edu.ec.



Oscar Moncayo Carreño Docente titular de la Carrera de Ingeniería en Gestión Empresarial de la Facultad de Ciencias Empresariales de la Universidad Técnica Estatal de Quevedo. Magister en Dirección de Empresas con Énfasis en Gerencia Estratégica en la Universidad Regional Autónoma de los Andes. Email omoncayo@uteq.edu.ec