A 'Phylogeny-aware' Multi-objective Optimization Approach for Computing MSA

ABSTRACT

Multiple sequence alignment (MSA) is a basic step in many analyses in bioinformatics, including predicting the structure and function of proteins, orthology prediction and estimating phylogenies. The objective of MSA is to infer the homology among the sequences of chosen species. Commonly, the MSAs are inferred by optimizing a single objective function. The alignments estimated under one criterion may be different to the alignments generated by other criteria, inferring discordant homologies and thus leading to different evolutionary histories relating the sequences. In the recent past, researchers have advocated for the multi-objective formulation of MSA, to address this issue, where multiple conflicting objective functions are being optimized simultaneously to generate a set of alignments. However, no theoretical or empirical justification with respect to a real-life application has been shown for a particular multi-objective formulation. In this study, we investigate the impact of multi-objective formulation in the context of phylogenetic tree estimation. In essence, we ask the question whether a phylogeny-aware metric can guide us in choosing appropriate multi-objective formulations. Employing evolutionary optimization, we demonstrate that trees estimated on the alignments generated by multi-objective formulation are substantially better than the trees estimated by the state-of-the-art MSA tools, including PASTA, T-Coffee, MAFFT etc.

CCS CONCEPTS

• **Applied computing** → **Bioinformatics**; *Multi-criterion optimization and decision-making*; • **Theory of computation** → *Evolutionary algorithms*;

KEYWORDS

Multiple sequence alignment, Phylogenetic tree, Evolutionary Multi-objective optimization

ACM Reference Format:

. 2019. A 'Phylogeny-aware' Multi-objective Optimization Approach for Computing MSA . In *Proceedings of the Genetic and Evolutionary Computation Conference 2019 (GECCO '19).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In biological research, multiple sequence alignment (MSA) is a useful and/or essential task in various applications such as phylogeny estimation, prediction of the structure and function of an RNA or protein, identification of functionally important sites, orthologous gene identification etc. The MSA task seeks to arrange more than two biological sequences based on certain criteria (such as evolutionary history, 3D structure etc.) by inserting spaces between letters in the sequences. In this research, we limit our focus on MSA in the context of phylogeny. Phylogeny estimation from molecular sequences generally operates as a two-phase approach. At first, the given sequences are aligned using an MSA method, and then a tree is estimated from the resultant alignment. The quality of inferred trees heavily depends on the quality of the corresponding alignment. Therefore, it is important to select an MSA tool that is the 'most suitable' in the phylogenetic context.

In this study, we make an attempt to identify a multi-objective formulation of MSA that is more effective in phylogeny estimation. Our motivation for a multi-objective formulation comes from the fact that the alignment estimated under one objective may be different to the alignments generated by other objectives, inferring discordant homologies and thus leading to different and often conflicting evolutionary histories relating the sequences under consideration. Multi-objective formulations can address this issue by optimizing multiple conflicting objectives simultaneously to generate a set of alignments. However, we are faced with the challenge of using appropriate measures/metrics to choose from among a number of objective sets to optimize. So, we ask the natural question whether the popular general purpose measures to judge the alignment quality can truly reflect the quality in the context of a particular application domain, i.e., phylogeny estimation in our case. While this question has received some shallow discussion in several studies [15, 18, 33], to the best of our knowledge no systematic investigation has been reported in the literature to this end. Therefore, in essence, we systematically investigate whether a phylogeny-aware metric can guide us better in choosing appropriate multi-objective formulation or tools capable of generating alignments that can produce better phylogenetic trees.

There are numerous tools available in the literature to compute MSA. We can broadly divide them into three groups: progressive techniques, consistency-based techniques and iterative techniques. This division is not exclusive as many tools also use a combination of these techniques. Progressive technique is the foundation of many MSA tools such as, Clustal Ω [27], PRANK [17], Kalign [14], FSA [2], RetAlign [29] etc. They compute the alignment using a guide tree by aligning pairs of sequences in a "bottom-up" manner. The consistency based techniques first construct a database of local and global pairwise alignments to facilitate generating an overall accurate alignment. The representatives of this category are T-Coffee [20], ProbCons [7], MSAProb [16], ProbAlign [23] etc. On the other hand, the iterative techniques were designed to achieve reliable alignments. These techniques try to fix the effect of mistakes made during the initial phases by repeating some crucial

GECCO '19, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

steps until some criteria are met. We find several examples of such techniques, such as, MAFFT [13], MUSCLE [8], MUMMALS [22], ProbConsetc. In this category, we also see some "meta-methods" such as, SATé [15] and PASTA [18], which co-estimate alignment and tree using other methods. These tools achieve scalability by employing the divide-and-conquer principle and are being used widely in practice.

The performance of an MSA tool is usually evaluated by comparing its output alignment with the reference alignment (provided with the dataset as the ground truth) in terms of several measures. To this end, the most popular measures are perhaps sum-of-pair (SP) score and total-column (TC) score. SP score is the fraction of the homologies (i.e., pairs of aligned characters) in the reference alignments recovered in the estimated alignment. Similarly, TC score is the fraction of the actual aligned columns that appear in the estimated alignment.

In this post-genomic era, the MSA datasets are posing new challenges to the researchers. Usually, an MSA method is provided with a default parameter configuration for aligning any problem instance with satisfactory accuracy. But these default values can not guarantee the best output throughout all kinds of datasets [24]. For instance, there is a parameter in ProbCons called the number of iterative refinement passes. Although it can vary between 0 to 1000 the default value is set to 100. We can achieve better results by tuning the parameter values which is not a straightforward task. Moreover, despite rigorous parameter tuning, no method can consistently outperform other methods for all datasets.

Therefore, we see the emergence of novel approaches that combine different alignment tools [32]. One such approach is evolutionary optimization where alignments generated from different tools are exploited to produce improved alignments without vesting any effort in parameter tuning. The success of such approach depends on the selection of proper objective function that can push the solutions (i.e., alignments) towards the desired zone that reflects the actual purpose of an alignment task. As any single objective function alone cannot be effective to tackle different challenges, it is wise to simultaneously optimize multiple objective functions. This will produce a set of competing solutions as the final output, which can be expected to contain our desired solution(s). Thus the evolutionary multi-objective optimization of MSA turns out to be appealing.

During the last decade, we find several studies [1, 3, 21, 25, 28, 34] with multi-objective formulations for MSA have been published – proposing two to four objective functions to capture and quantify different aspects of an alignment. Among them, probably the most popular is the sum-of-pairs score and its weighted variants, where a pairwise score is calculated for each pair of aligned sequences using a substitution matrix. This matrix should reflect the characteristics of the data at hand. Although we know that the same character across all rows of a column does not necessarily indicate homology, the count of such columns in an alignment is seen as a maximization objective known as totally conserved columns. Next, we find attempts to minimize the total number of gaps to maintain the compactness of an alignment. Then there are different types of gap penalties that penalize each sequence for introducing gaps. Also, we find two other objective functions, Entropy and Similarity, that compute column-wise scores and then sum those together. Both of them try to express the homogeneity of characters in a column using two different ways. Contrary to the performance/quality measures mentioned earlier (such as SP score, TC score), we are not allowed to use the reference alignment while calculating these objective functions.

We notice several issues in the works advocating multi-objective formulation of MSA (in the context of different applications where the MSA will be used). First of all, in these works, there is a lack of sound theoretical or empirical justification for the choice of a particular objective function to be optimized. Secondly, we also notice the absence of a sound rationale/justification behind the two most popular performance metrics, namely, sum-of-pair score and total column score. On the contrary, it seems only natural that performance score should reflect the actual purpose of MSA. For example, if the goal is to estimate a phylogenetic tree, the performance metric to be used for evaluation should be able to accurately measure the quality and usefulness of the constructed tree. Notably, another issue, specific to the domain of phylogeny estimation, is the use of relatively smaller (number of taxa below 50) datasets in experiments.

In this article, we attempt to demonstrate the effectiveness of multi-objective MSA by addressing the above mentioned limitations in the context of its intended application domain (i.e., phylogeny estimation). To make a fair comparison with nine state-of-the-art MSA tools, we conduct comprehensive experimentations on both simulated and biological datasets using tree as well as alignment quality measures. In particular, this article makes the following key contributions:

- To the best of our knowledge, this is the first attempt to investigate the effect of using domain-specific measures (as opposed to generic alignment measures) to evaluate the performance of MSA methods in the context of phylogeny estimation.
- We suggested a methodology based on multiple linear regression to judge the potential efficacy of a multi-objective formulation of MSA. Then, based on this methodology, we identified a multi-objective formulation that had the potential to yield better phylogenetic trees.
- Finally, we demonstrated that the multi-objective formulation can consistently yield better phylogenetic trees than several state-of-the-art MSA tools. And interestingly we found that popular alignment quality measures do not necessarily lead to highly accurate phylogenetic trees.

2 METHODS

We begin this section with an overview of our experimental design. Then we present the multi-objective formulations selected to compute MSA. Finally, we discuss the multi-objective evolutionary algorithms applied to optimize those objective functions as well as the state-of-the-art tools that we utilized in this study.

2.1 Experimental design

Our experimental methodology is briefly described below (please see also Figure 1):

- Step 1: Following a systematic approach involving multiple linear regression applied on a simulated dataset, we first

make an attempt to identify and choose one multi-objective formulation that turns out to be potentially more effective in the context of phylogeny estimation (discussed in Section S6 of the supplementary file).

- Step 2: We run a popular and effective multi-objective evolutionary algorithm on biological datasets to optimize the set of objective functions selected in Step 1. Each run of the evolutionary algorithm on each dataset gives us a set of alignments as output.
- Step 3: We also run nine state-of-the-art MSA tools (please see Table 3) to generate alignments on all these datasets.
- Step 4: We evaluate the quality of each generated alignment with respect to the reference alignment using two popular measures, namely, SP score and TC score (discussed in Section S3 of the supplementary file).
- Step 5: For each of the generated alignments, we infer maximum likelihood (ML) phylogenetic tree (discussed in Section S4 of the supplementary file). Then we measure the quality of each inferred tree with respect to the reference tree (true tree) using the mostly used measure in the literature called false negative (FN) rate (discussed in Section S5 of the supplementary file).
- Step 6: Finally we compare the alignments and the corresponding ML trees generated by the multi-objective optimization with the ones generated by the state-of-the-art tools.

2.2 Multi-objective formulations

A multi-objective formulation defines the problem using a set of objective functions to be optimized simultaneously. In this study, we have selected the following three multi-objective formulations of MSA from the literature based on their simplicity as well as performance as reported in the literature.

- {SOP, TC}: Maximize the sum of pairs (SOP) and the number of totally aligned columns (TC) [3].
- {Gap, SOP}: Maximize the sum of pairs (SOP) and minimize the number of gaps (Gap) [1].
- {wSOP, TC}: Maximize the weighted sum of pairs with affine gap penalties (wSOP) and the number totally aligned columns (TC) [25].

We describe these objective functions along with several existing ones in Section S1 of the supplementary file. From onward, we use the terms shown in Table 1 to refer to these objective functions.

Table 1: Terms used to denote objective functions.

#	Objective function	Term
1	Maximize no. of totally aligned columns	TC
2	Minimize no. of gaps	Gap
3	Maximize sum of pairs	SOP
4	Maximize weighted sum of pairs with affine gap penalties	wSOP

We need a substitution matrix to calculate SOP and wSOP. In this study, we used NUC4.4 (supplied by NCBI at ftp://ftp.ncbi.nih.gov/blast/matrices/NUC.4.4) for nucleotide sequences and BLOSUM62 [10] for protein sequences.

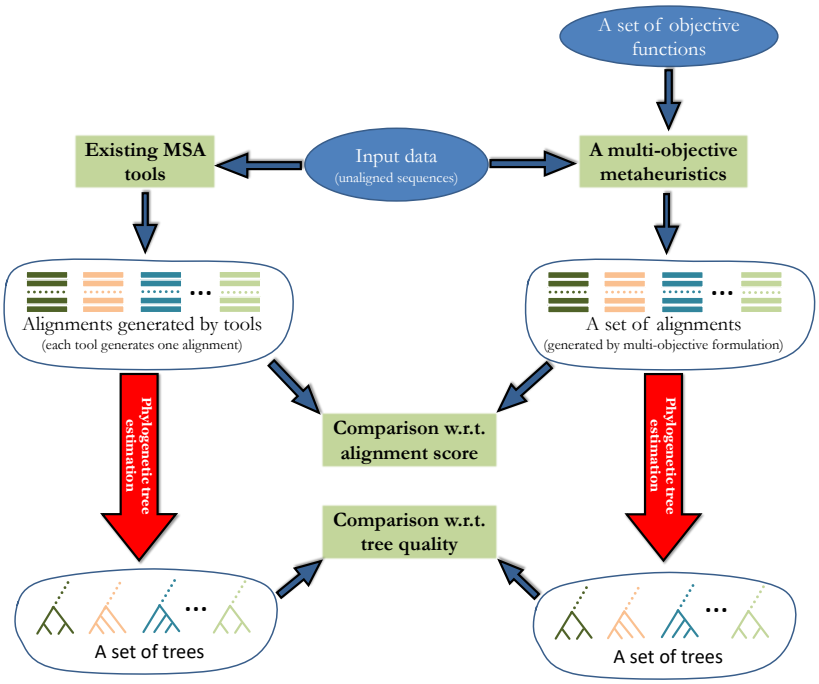


Figure 1: Our methodology for finding the impact of a multi-objective formulation (i.e., a set of objective functions) of MSA on phylogenetic tree estimation. For each dataset (i.e., unaligned sequences), we run a multi-objective evolutionary algorithm. It simultaneously optimizes the given objective functions and outputs a set of alignments which represents the best-possible compromise among all objective functions. We also run several existing MSA tools on that dataset and each tool generates one alignment. We evaluate the quality of each generated alignment with respect to the reference alignment using widely used scores. Also, we estimate phylogenetic trees for all alignments and evaluate each tree with respect to the reference. Then we compare the alignments and the corresponding phylogenetic trees generated by the multi-objective formulation with the ones generated by the existing tools based on alignment score as well as tree quality. We also observe the association between alignment scores and tree quality values to examine whether it is appropriate to use alignment score in the context of phylogeny estimation.

2.3 Multi-objective Evolutionary Algorithms

To simultaneously optimize multiple objective functions, we ran two popular multi-objective evolutionary algorithms: NSGA-II [5] and NSGA-III [4]. Two studies ([21, 35]) demonstrated the strength of NSGA-II for computing MSA. NSGA-II works best when the number of objectives is upto three while NSGA-III is specially designed for handling more than three objectives. Hence, we applied these algorithms according to Table 2. We discuss these methods along with their vital components and parameters in Section S2 of the supplementary file. We implemented them using jMetalMSA [36] which is a Java metaheuristic framework for MSA.

Table 2: Our selected algorithms and corresponding objective set.

Algorithm	Objective set
NSGA-II	{Gap, SOP}, {SOP, TC}, {wSOP, TC}
NSGA-III	{Gap, SOP, wSOP, TC}

2.4 State-of-the-art MSA tools

We used the alignments generated by nine representative state-of-the-art MSA tools (shown in Table 3) to compare with our approach. We run each of them with its default parameter configuration. Moreover, we initialize the multi-objective evolutionary algorithms with a set of alignments generated by randomly mixing and modifying those nine alignments. Notably, this approach, known as the seeded initial population generation, is quite common in the metaheuristics literature specially for multi-objective optimization.

Table 3: List of state-of-the-art MSA tools that we used in this study.

For nucleotide sequences		For protein sequences	
Tool	Version	Tool	Version
FSA [2]	1.15.9	FSA	1.15.9
PASTA [18]	1.7.8	PASTA	1.7.8
T-Coffee [20]	11.00	T-Coffee	11.00
MAFFT [13]	7.31	MAFFT	7.245
Clustal W [30]	2.1	Clustal W	2.1
Clustal Ω [27]	1.2.4	RetAlign [29]	1.0
MUSCLE [8]	3.8.31	MUSCLE	3.8.31
PRANK [17]	0.170427	ProbCons [7]	1.12
Kalign [14]	2.03	Kalign	2.04

3 RESULTS

We conducted extensive experiments with both simulated and biological datasets. We begin by carefully and systematically selecting a multi-objective formulation which is potentially useful for phylogenetic tree estimation employing NSGA-III and multiple linear regression. Next, we generate alignments through running NSGA-II as well as nine state-of-the-art MSA tools. Then we compare those alignments with respect to both generic and domain-specific quality measures. In what follows, unless otherwise specified, when we discuss the (best) results of a tool, we mean one of the above-mentioned nine tools.

3.1 Datasets

We studied the 100-taxon simulated dataset [15] and the widely used BAliBASE 3.0 benchmark [31] which is a biological dataset. As the simulated dataset comes with the true phylogenetic tree, we use this dataset to examine whether a multi-objective formulation of MSA is potentially phylogeny-aware and in the sequel, we select one such formulation somewhat similar to the training phase of a machine learning approach. Afterward, we validate the effectiveness of the selected formulation against the state-of-the-art MSA tools based on the biological dataset.

We randomly selected five replicates from the 100-taxon simulated dataset and 27 instances from the BAliBASE 3.0 benchmark. Section S7.1 of the supplementary file provides a detailed description of these datasets.

3.2 Selection of an appropriate multi-objective formulation

As has been mentioned above, we have used the 100-taxon simulated dataset to select one multi-objective formulation that has the potential to be 'phylogeny-aware'. To reduce the computational effort, we pre-select three multi-objective formulations of MSA and limit our investigation thereon. Thus we choose one of the formulations from among {Gap, SOP} [1], {SOP, TC} [3] and {wSOP, TC} [25] (please see Section 2.2 Table 1). We experiment with five randomly selected replicates (R0, R4, R9, R14, R19) and then judge based on two criteria: firstly, we used multiple linear regression analysis to examine the association between individual objective function and FN rate; secondly, we assess the alignments generated through the optimization of each set of objective functions in terms of resultant ML trees.

We need to consider the relationship between each pair of objective functions to properly interpret the result of multiple linear regression. We perform this by running NSGA-III [4] for 25 times which optimizes all the objective functions (i.e., {Gap, SOP, wSOP, TC}) and thus we obtain a large collection of diverse alignments. A visualization of the interrelations among the objective values of those solutions is presented in Figure S2 of the supplementary file. From these experiments, we have the following two key observations.

(a) In all the cases, SOP is totally correlated with wSOP. So we do not need to optimize both of them. Moreover, this high correlation creates a serious problem in multiple regression analysis called multicollinearity [19]. Therefore, we should not keep these two objective functions together in our regression analysis. Also, it is redundant to consider both of them in the multi-objective formulation.
(b) SOP is clearly in conflict with Gap across all the replicates. Therefore, if we optimize them simultaneously, we can generate many diverse solutions which represent the compromise between these two objective functions [12]. This diverse collection is likely to contain the desired alignment for any kind of dataset.

As the objective functions are inter-related, we need to measure the degree of association between an objective and FN rate while holding the remaining objectives constant to avoid getting any spurious result [19]. Therefore, we perform multiple linear regression by employing the following model:

$$\text{FN rate} = \beta_0 + \beta_1 \times \text{TC} + \beta_2 \times \text{Gap} + \beta_3 \times \text{SOP (or wSOP)} + \epsilon \quad (1)$$

Each coefficient (β_1, β_2 and β_3) represents the expected change in the FN rate per unit change in the corresponding objective function when all the remaining objective functions are held constant. For this reason, they (β_i) are called partial regression coefficients. ϵ is the random error component which is assumed to follow a Gaussian distribution with mean zero and some fixed standard deviation. We fit this model to the solutions generated by optimizing the set {Gap, SOP, wSOP, TC}. For each of those solutions, we estimate ML tree and evaluate its quality in terms of FN rate. We estimate

these coefficients using the least-squares method (an illustration is presented in Figure S3 of the supplementary file). We apply t-test on individual regression coefficient (i.e., slope) β_i (with null hypothesis $\beta_i = 0$) to test the significance of that association. We can note the following two interesting points from these results.

(a) In the majority of the cases (R0, R4 and R14), Gap, SOP and wSOP exhibit a good degree of association with FN rate (i.e positive slope) with high confidence (p-value close to 0) compared to other objective functions. So, we can expect them to be good optimization objectives for MSA.
(b) For replicate R4 and R19, none of the objective exhibit good association. This shows that an objective function might not perform well across all problem instances.

Now we measure the strength of each objective set based on the FN rate achieved by the members of the generated solution set. To accomplish this, For each set of objective functions, we run NSGA-II [5] for 20 times following the standard practice of operations research (OR) literature (due to the stochastic nature of metaheuristics). Each run generates a set of solutions that represent the trade-offs in satisfying all objectives. Afterward, we inferred ML tree for each of the generated alignment. We collected the best FN rates from each of the 20 solution sets and examine the distribution of these FN rates (a visualization of these distributions using boxplots is presented in Figure S4 of the supplementary file). Here we have the following key observations:

- For most of the cases, the combined set {TC, Gap, SOP, wSOP} achieves better results than the other sets. This indicates that adding suitable objective functions increase the chance of achieving the best FN rate. However, this increases the overall complexity of the multi-objective evolutionary algorithm. So in this study, we keep the size of the objective set as small as possible.
- Among our three pre-selected objective sets, {Gap, SOP} achieves relatively lower FN rates. This is consistent with our regression results discussed earlier.
- Both {TC, Gap, SOP, wSOP} and {Gap, SOP} persistently generate better FN rates than the state-of-the-art tools.

Based on our findings discussed so far, we consider {Gap, SOP} to be the most suitable candidate to conduct our study among all the formulations considered above.

3.3 Validation of the selected multi-objective formulation

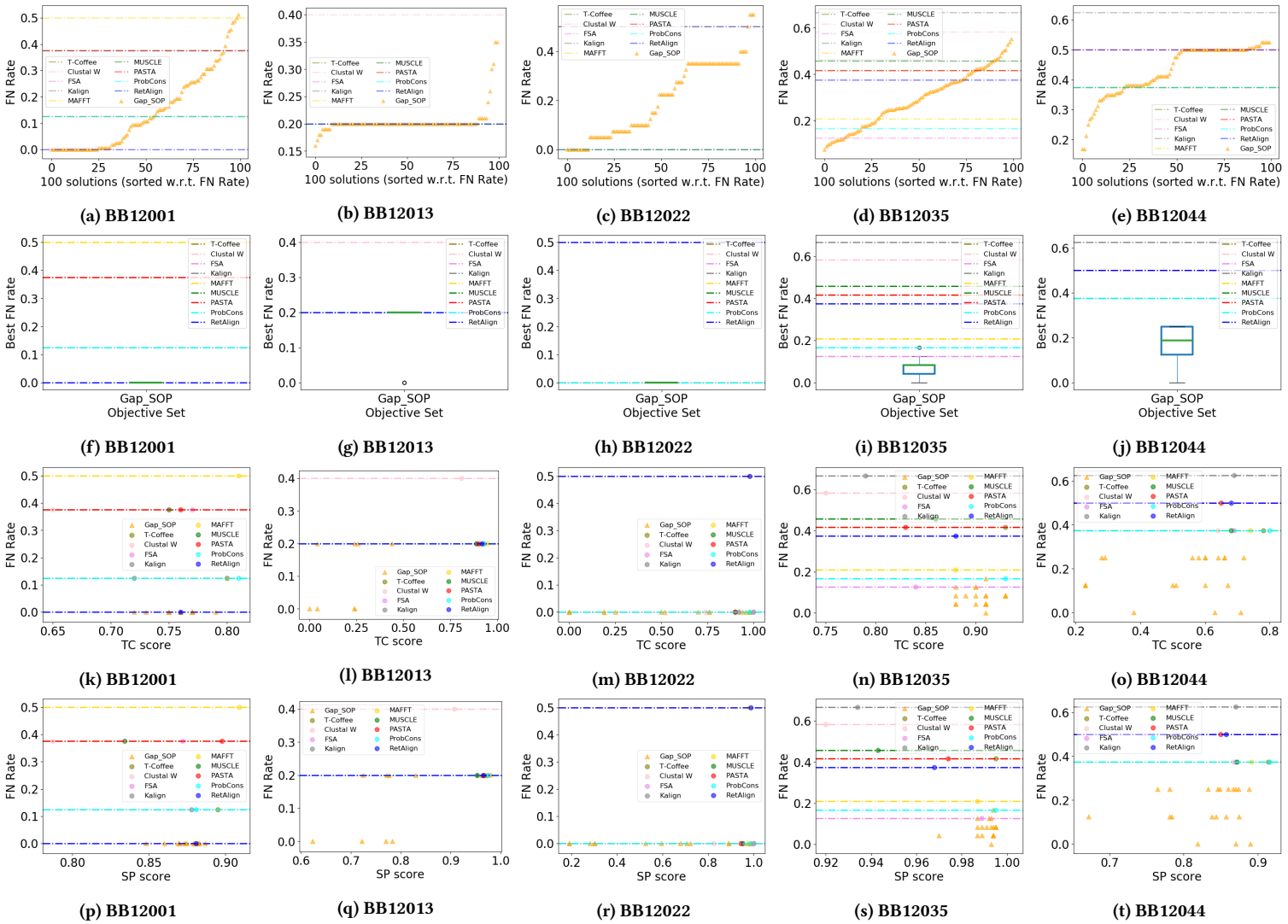
To judge the effectiveness of our chosen formulation (i.e. {Gap, SOP}), we conducted 20 independent runs of NSGA-II for each of the randomly selected BAliBASE datasets under six groups (RV11, RV12, RV20, RV30, RV40 and RV50). Here we analyze the generated solutions based on the quality of alignments and resultant trees. We witnessed that the alignments which are better according to the widely accepted alignment scores, not necessarily generate better phylogenetic trees. Here we discuss our key observations on the selected four datasets (BB12001, BB12013, BB12022, BB12035 and BB12044) under the group RV12 using Figures 2 and 3. For the remaining groups (RV11, RV20, RV30, RV40 and RV50), our core findings are similar and consistent. For the sake of brevity, we present those results in Section S7 of the supplementary file.

In Panel 1 and 2 of Figure 2, we compare the performance of {Gap, SOP} with respect to FN rate against the nine state-of-the-art tools from two perspectives. Here, part (a) - (e) show the averaged FN rate of 100 solutions over 20 runs. Since each run generates 100 solutions, we make the average meaningful by sorting the 100 FN rates per run. Then we average the best FN rates across all the runs. The same applies to the second best ones and so on. And part (f) - (j) summarize the variation of the best FN rate (among 100 values) across 20 runs. We see that {Gap, SOP} outperforms all the state-of-the-art tools for BB12013, BB12035 and BB12044. In the case of BB12044, {Gap, SOP} reconstructs all the edges correctly as opposed to 37% FN rate attained by the tree estimated on the MSA generated by the best tool MUSCLE which is remarkable. For the remaining datasets (BB12001 and BB12022), {Gap, SOP} performs as good as the best tool. On all the datasets, {Gap, SOP} generates several solutions that are equivalent or better than that of the best tool.

We perform a similar analysis based on the widely used two alignment quality measures, namely, TC score and SP score and report the results in Figure 3. With respect to TC score (part (a) - (e) of Figure 3), {Gap, SOP} can outperform all the tools only for BB12001 which is contrary to the findings based on FN rate. So we see the disagreement between FN rate and TC score which we examine graphically in part (k) - (o) of Figure 2. If we observe the results based on SP score (part (k) - (o) of Figure 3), we get similar disagreement between FN rate and SP score which is illustrated in part (p) - (t) of Figure 2. We find that there are several solutions that achieve better FN rates in spite of their poor alignment quality (TC and SP score). We consistently observe this phenomenon across the remaining datasets as well which we present in Section S7 of the supplementary file. From this analysis, we realize that the tools/approaches achieving better performance than {Gap, SOP} in terms of the popular measures, namely, TC score and SP score fail to achieve better FN rates than {Gap, SOP}. Even from among the tools, there is disagreement between alignment quality score and FN rate.

Table 4 shows a comparative summary of the 100 solutions generated by a single run of NSGA-II while optimizing {Gap, SOP} with respect to the nine state-of-the-art MSA tools based on FN rate for the 27 randomly selected BAliBASE datasets. Here we see that, the multi-objective formulation has been able to generate better phylogenetic trees than all the state-of-the-art MSA tools except on a few cases (marked by cells with 0 value).

3.3.1 Statistical significance. Now we confirm the significance of the improvement achieved by NSGA-II while optimizing {Gap, SOP} (denoted here as NSGA-II$_{\{Gap,\ SOP\}}$) in terms of FN rate over nine MSA tools on 27 BAliBASE datasets by applying an appropriate statistical test. We form paired data by picking the FN rate achieved by each (MSA method, dataset) pair. For NSGA-II, we take the average of the 20 best FN rates from 20 independent runs considering its stochastic nature. As our data do not satisfy the condition of normality and homoscedasticity [26], we choose a series of non-parametric tests following the recommendation of [6]. At first, we simultaneously compare all the methods using the Friedman test [9]



(a) BB12001 (b) BB12013 (c) BB12022 (d) BB12035 (e) BB12044

(f) BB12001 (g) BB12013 (h) BB12022 (i) BB12035 (j) BB12044

(k) BB12001 (l) BB12013 (m) BB12022 (n) BB12035 (o) BB12044

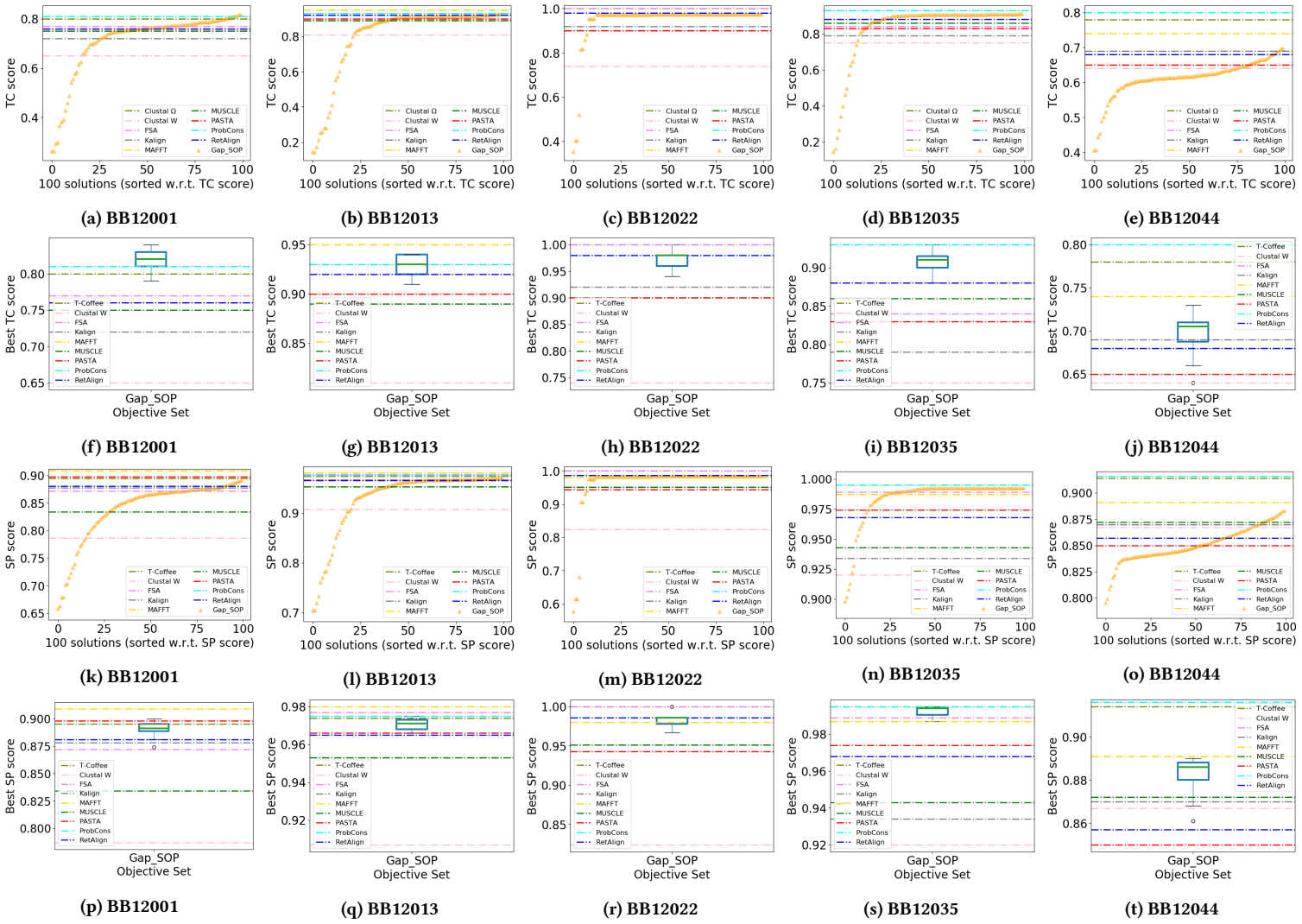
(p) BB12001 (q) BB12013 (r) BB12022 (s) BB12035 (t) BB12044

Figure 2: RV12: Panel 1 (Top panel): part (a) - (e) show the FN rate of 100 solutions averaged over 20 runs. At first, we sort the FN rates of each solution set. Then we average the FN rates at each sorted position of all the sets. Panel 2: part (f) - (j) show the distribution of the best FN rates collected from all runs. Panel 3 (Panel 4): part (k) - (o) (part (p) - (t)) show the relationship between FN rate and TC score (SP score) for different alignments. In all panels, we show the FN rates achieved by the nine state-of-the-art tools using dashed horizontal lines.

which gives the relative ranking (lower is better) of all the methods and strongly suggests the existence of significant differences among the methods considered (as p-value is 0). The results have been presented in Column 2 of Table 5. Here we see that the multi-objective formulation achieve the top two positions. Next, we complement the Friedman test by following Holm's post-hoc procedure [11] to contrast the difference between the multi-objective formulation and each of the nine tools. The results have been summarized in Columns 3 of Table 5. Here, each cell shows the adjusted p-value which indicates the significance of difference in performance (based on FN rate) between two methods. We notice that all the p-values are very close to 0. So we can state with high confidence that, our chosen multi-objective formulation achieves statistically significant improvement over the nine MSA tools.

4 DISCUSSION & CONCLUSION

In this study, we have introduced a phylogeny-aware multi-objective optimization approach to compute MSA with an ultimate goal to infer the phylogenetic tree from the resultant alignments. To optimize MSA, we proposed two simple objective functions in addition to the existing ones. We judged the potential capability of each objective function to yield better trees by employing domain knowledge as well as by applying statistical approaches. We employed multiple linear regression to measure the degree of association between the



(a) BB12001 (b) BB12013 (c) BB12022 (d) BB12035 (e) BB12044

(f) BB12001 (g) BB12013 (h) BB12022 (i) BB12035 (j) BB12044

(k) BB12001 (l) BB12013 (m) BB12022 (n) BB12035 (o) BB12044

(p) BB12001 (q) BB12013 (r) BB12022 (s) BB12035 (t) BB12044

Figure 3: RV12: Panel 1 (Panel 3): part (a) - (e) (part (k) - (o)) shows the TC score (SP score) of 100 solutions averaged over 20 runs. At first, we sort the TC scores (SP scores) of each solution set. Then we average the TC scores (SP scores) at each sorted position of all the sets. Panel 2 (Panel 4): part (f) - (j) (part (p) - (t)) shows the distribution of the best TC scores (SP scores) collected from all runs. In all panels, we show the performance of nine state-of-the-art tools using dashed horizontal lines.

individual objective functions and the quality of the inferred phylogenetic tree (i.e., FN rate). Thus, we provide empirical justification to choose a multi-objective formulations to move forward. Afterward, we performed extensive experimentation with biological datasets to demonstrate the benefit of our approach. We showed that the simultaneous optimization of a set of phylogeny-aware objective functions can lead to phylogenetic trees with improved accuracy than that of the state-of-the-art MSA tools. From this finding, we would like to hypothesize that, the use of domain-specific measures can aid MSA methods in other application domain as well.

Standard criteria (SP-score, TC-score, etc.) for assessing alignment quality are usually based on shared homology pairs (SP score) or identical columns (TC score), and do not explicitly consider a particular application domain. Mistakes in alignments that are not important with respect to an application domain may not impact the ultimate accuracy of that particular inference. For example, not all sites are significant with respect to protein structure and function prediction, and hence multiple alignments with different accuracy may lead to the same predictions [33]. Similarly, in the context of phylogeny estimation, alignments with substantially different SP scores may lead to trees with the same accuracy [15]. In this study, we systematically investigate the impact of evaluation criteria of an alignment on phylogenetic tree inference problem. Our results suggest that it could be possible to develop improved MSA methods for phylogenetic analysis by carefully choosing appropriate objective functions. Moreover, in almost all existing studies on MSA, we find the researchers evaluating the effectiveness of MSA methods using some generic alignment quality measures (i.e., TC score, SP score).

Table 4: Comparative summary of the 100 solutions generated by a single run of NSGA-II while optimizing {Gap, SOP} with respect to the nine state-of-the-art MSA tools based on FN rate.

Group	Dataset	Avg. no. of solutions (out of 100) generated by a single run of NSGA-II which are **better or equivalent** to a MSA tool according to FN rate								
		T-Coffee	Clustal W	FSA	Kalign	MAFFT	MUSCLE	PASTA	ProbCons	RetAlign
RV11	BB11005	100	2	100	100	54	54	86	86	86
	BB11018	36	64	100	86	36	86	97	17	17
	BB11033	71	71	97	71	71	0	9	71	9
	BB11020	34	34	100	100	34	0	0	0	34
RV12	BB12001	55	92	92	55	99	92	92	55	25
	BB12013	9	100	9	9	9	9	9	9	9
	BB12022	12	12	12	12	97	12	12	12	97
	BB12035	80	100	11	100	29	90	80	20	73
	BB12044	23	23	23	100	23	23	88	23	88
RV20	BB20001	92	0	1	0	0	92	0	0	15
	BB20010	23	99	1	7	77	23	77	23	7
	BB20022	82	59	82	100	82	82	59	59	0
	BB20033	96	5	96	19	82	29	68	58	88
	BB20041	57	71	38	22	30	83	65	51	71
RV30	BB30002	0	85	45	7	45	0	7	7	7
	BB30008	53	53	98	25	98	98	100	38	90
	BB30015	61	93	93	88	88	61	100	88	88
	BB30022	64	19	47	0	2	47	19	5	84
RV40	BB40001	60	86	60	41	75	41	97	60	97
	BB40013	51	39	39	45	26	45	45	62	62
	BB40025	0	0	0	0	0	49	49	49	49
	BB40038	26	90	66	15	15	66	0	2	66
	BB40048	69	89	69	89	69	89	69	69	69
RV50	BB50001	10	85	62	94	10	100	62	85	85
	BB50005	0	93	0	93	93	93	93	0	93
	BB50010	0	0	0	0	0	28	0	0	96
	BB50016	66	96	0	78	66	78	54	96	78

Contrastingly, our results revealed that optimizing those widely used measures do not necessarily lead us to the best phylogenetic tree. This finding could be an eye-opener for the researchers who need to use MSA methods to address a particular application.

Our findings and proposed multi-objective formulation can be particularly beneficial for iterative methods like SATé and PASTA that iteratively co-estimate both alignment and tree. These methods obtain an initial alignment and a tree that guide each other to improved estimates in an iterative fashion. They make an effort to exploit the close association between the accuracy of an MSA and the corresponding tree in finding the output through multiple iterations from both directions. Therefore, carefully choosing an evaluation metric for an MSA with better correlation to the tree accuracy seems likely to improve the results of these co-estimation techniques. Thus, our methodology, if adopted, may potentially have a profound positive impact on the accuracy of these iterative co-estimation techniques.

This study will encourage the scientific community to investigate various application-aware measures for computing and evaluating MSAs. This will potentially prompt more experimental studies addressing specific application domains; and ultimately will propel our understanding of MSAs and their impact in various domains in computational biology, i.e, phylogeny estimation, protein structure and function prediction, orthology prediction etc. This study will also encourage the researchers to develop new scalable MSA tools by simultaneously optimizing multiple appropriate optimization criteria. Thus, we believe that this study will pioneer new models and optimization criteria for computing MSA – laying a firm, broad foundation for application-specific multi-objective formulation for estimating multiple sequence alignment.

Table 5: Friedman test (Column 2): The Average Friedman's ranking (lower is better) achieved by the MSA methods over 27 BAliBASE datasets. We performed the Friedman test based on FN rate achieved by the tools. For NSGA-II, we consider the average of the 20 best FN rates obtained from 20 runs. We also show the computed statistics and corresponding p-value. Holm's post-hoc procedure (Columns 3): Comparison between NSGA-II and the MSA tools using the Holm's post-hoc procedures (as a complement of the Friedman test) over 27 BAliBASE datasets. Each entry shows the adjusted p-value which indicates the significance of difference in performance (based on FN rate) between two methods.

1	2	3
Method	Friedman's Rank*	Holm's adjusted p-value NSGA-II$_{\{Gap, SOP\}}$
NSGA-II$_{\{Gap, SOP\}}$	2.8704	-
ProbCons	5.7963	0.00238
Clustal Ω	6.2963	0.00044
MAFFT	6.4074	0.00036
Kalign	6.7037	0.00011
PASTA	6.8148	0.00007
FSA	6.9444	0.00004
MUSCLE	7.1482	0.00002
Clustal W	7.3519	0.00001
RetAlign	7.4630	0.00000
*Statistic	10.5911	N/A
*p-value	0.00000	

We performed an extensive experimental study comprising 27 datasets of varying sizes and complexities, and our findings are consistent throughout all the datasets. Still, we acknowledge the possibility of facing a few unforeseen circumstances as follows. There might be some datasets on which our approach might not exhibit satisfactory performance. Besides, currently we did not pay any effort to improve the running time of our approach which is higher as compared to top MSA tools. However, sufficient speedup could be achieved by leveraging the modern computing architectures (computer cluster, GPU, etc.).

Formulating phylogeny-aware multi-objective formulation (application specific evaluation criteria in general) cannot be developed entirely in one study; it should evolve in response to scientific findings and systematists' feedback. This requires the active involvement of evolutionary biologists, computer scientists, systematists, and others – leading to improved understandings of alignments and how they are related to various fields in comparative genomics.

REFERENCES

[1] Maryam Abbasi, Luís Paquete, and Francisco B Pereira. 2015. Local search for multiobjective multiple sequence alignment. In *International Conference on Bioinformatics and Biomedical Engineering*. Springer, 175–182.

[2] Robert K Bradley, Adam Roberts, Michael Smoot, Sudeep Juvekar, Jaeyoung Do, Colin Dewey, Ian Holmes, and Lior Pachter. 2009. Fast statistical alignment. *PLoS computational biology* 5, 5 (2009), e1000392.

[3] Fernando José Mateus da Silva, Juan Manuel Sánchez Pérez, Juan Antonio Gómez Pulido, and Miguel A Vega Rodríguez. 2010. AlineaGA—a genetic algorithm with local search optimization for multiple sequence alignment. *Applied Intelligence* 32, 2 (2010), 164–172.

[4] Kalyanmoy Deb and Himanshu Jain. 2014. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: Solving problems with box constraints. *IEEE Trans. Evolutionary Computation* 18, 4 (2014), 577–601.

[5] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation* 6, 2 (2002), 182–197.

[6] Joaquín Derrac, Salvador García, Daniel Molina, and Francisco Herrera. 2011. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation* 1, 1 (2011), 3–18.

[7] Chuong B Do, Mahathi SP Mahabhashyam, Michael Brudno, and Serafim Batzoglou. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome research* 15, 2 (2005), 330–340.

[8] Robert C Edgar. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32, 5 (2004), 1792–1797.

[9] Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association* 32, 200 (1937), 675–701.

[10] Steven Henikoff and Jorja G Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* 89, 22 (1992), 10915–10919.

[11] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.

[12] Deb Kalyanmoy. 2001. *Multi objective optimization using evolutionary algorithms*. John Wiley and Sons.

[13] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* 30, 14 (2002), 3059–3066.

[14] Timo Lassmann, Oliver Frings, and Erik LL Sonnhammer. 2008. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic acids research* 37, 3 (2008), 858–865.

[15] Kevin Liu, Sindhu Raghavan, Serita Nelesen, C Randal Linder, and Tandy Warnow. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* 324, 5934 (2009), 1561–1564.

[16] Yongchao Liu, Bertil Schmidt, and Douglas L Maskell. 2010. MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics* 26, 16 (2010), 1958–1964.

[17] Ari Löytynoja and Nick Goldman. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National academy of sciences of the United States of America* 102, 30 (2005), 10557–10562.

[18] Siavash Mirarab, Nam Nguyen, Sheng Guo, Li-San Wang, Junhyong Kim, and Tandy Warnow. 2015. PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology* 22, 5 (2015), 377–386.

[19] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. 2012. *Introduction to linear regression analysis*. Vol. 821. John Wiley & Sons.

[20] Cédric Notredame, Desmond G Higgins, and Jaap Heringa. 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment1. *Journal of molecular biology* 302, 1 (2000), 205–217.

[21] Francisco M Ortuño, Olga Valenzuela, Fernando Rojas, Hector Pomares, Javier P Florido, Jose M Urquiza, and Ignacio Rojas. 2013. Optimizing multiple sequence alignments using a genetic algorithm based on three objectives: structural information, non-gaps percentage and totally conserved columns. *Bioinformatics* 29, 17 (2013), 2112–2121.

[22] Jimin Pei and Nick V Grishin. 2006. MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information. *Nucleic acids research* 34, 16 (2006), 4364–4374.

[23] Usman Roshan and Dennis R Livesay. 2006. Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics* 22, 22 (2006), 2715–2721.

[24] Álvaro Rubio-Largo, Leonardo Vanneschi, Mauro Castelli, and Miguel A Vega-Rodríguez. 2018. A characteristic-based framework for multiple sequence aligners. *IEEE transactions on cybernetics* 48, 1 (2018), 41–51.

[25] Álvaro Rubio-Largo, Miguel A Vega-Rodríguez, and David L González-Álvarez. 2016. A hybrid multiobjective memetic metaheuristic for multiple sequence alignment. *IEEE Transactions on Evolutionary Computation* 20, 4 (2016), 499–514.

[26] David J Sheskin. 2003. *Handbook of parametric and nonparametric statistical procedures*. crc Press.

[27] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* 7, 1 (2011), 539.

[28] Wilson Soto and David Becerra. 2014. A multi-objective evolutionary algorithm for improving multiple sequence alignments. In *Brazilian Symposium on Bioinformatics*. Springer, 73–82.

[29] Adrienn Szabó, Ádám Novák, István Miklós, and Jotun Hein. 2010. Reticular alignment: A progressive corner-cutting method for multiple sequence alignment. *BMC bioinformatics* 11, 1 (2010), 570.

[30] Julie D Thompson, Desmond G Higgins, and Toby J Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* 22, 22 (1994), 4673–4680.

[31] Julie D Thompson, Patrice Koehl, Raymond Ripp, and Olivier Poch. 2005. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins: Structure, Function, and Bioinformatics* 61, 1 (2005), 127–136.

[32] Julie D Thompson, Benjamin Linard, Odile Lecompte, and Olivier Poch. 2011. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PloS one* 6, 3 (2011), e18093.

[33] Tandy Warnow. 2013. Large-scale multiple sequence alignment and phylogeny estimation. In *Models and algorithms for genome evolution*. Springer, 85–146.

[34] Cristian Zambrano-Vega, Antonio J Nebro, José García-Nieto, and José F Aldana-Montes. 2017. Comparing multi-objective metaheuristics for solving a three-objective formulation of multiple sequence alignment. *Progress in Artificial Intelligence* (2017), 1–16.

[35] Cristian Zambrano-Vega, Antonio J Nebro, José García-Nieto, and Jose F Aldana-Montes. 2017. M2Align: parallel multiple sequence alignment with a multi-objective metaheuristic. *Bioinformatics* 33, 19 (2017), 3011–3017.

[36] Cristian Zambrano-Vega, Antonio J Nebro, José García-Nieto, and José F Aldana-Montes. 2017. A Multi-objective Optimization Framework for Multiple Sequence Alignment with Metaheuristics. In *International Conference on Bioinformatics and Biomedical Engineering*. Springer, 245–256.