Advanced Studies in Biology, Vol. 8, 2016, no. 1, 9 - 16
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/asb.2016.51250

Optimizing Genetic Algorithm Parameters for Multiple Sequence Alignment Based on Structural Information

May Rashiele K. Sueño[1]

Department of Mathematics and Computer Science
University of the Philippines
Baguio, Baguio, Philippines

Joel M. Addawe

Department of Mathematics and Computer Science
University of the Philippines
Baguio, Baguio, Philippines

Abstract

Multiple sequence alignments (MSAs) are commonly used approaches in the analysis of sequence structure relationships. MSA is generally the alignment of three or more protein or nucleic acid sequences that maximises the similarities between sequences.

In this paper, we use genetic algorithm to compute multiple sequence alignment using the structural information as the scoring scheme implemented in the program Multiobjective Optimizer for Sequence Alignments based on Structural Evaluations (MOSAStrE). We performed numerical experiments on datasets obtained from benchmark alignment database (BAliBASE) to solve multiple sequence alignment. To test the performance of the proposed algorithm, numerical simulations were carried out in deciding the appropriate set of parameter values for the

[1]Corresponding author

proposed algorithm. The results obtained are reported and discussed in this paper.

Keywords: Genetic Algorithm (GA), Multiple sequence alignments, scoring scheme

1 Introduction

Multiple sequence alignment (MSA) is generally the alignment of three or more protein or nucleic acid sequences that maximizes the similarities between them. MSAs are used in protein structure modeling, functional prediction and phylogenetic analysis, such as in the studies of [2, 4], which enables us to determine the evolutionary relationships between the sequences being studied. Many MSA methodologies have used different algorithms in their implementations including progressive and iterative approaches. Progressive algorithms work by creating a succession of pairwise alignments between the sequences to be aligned while iterative algorithms improve a mutiple sequence alignment by aligning it over successive iterations. Among the algorithms mentioned, progressive alignments are by far the most popular heuristic strategies with ClustalW being the most widely used implementation [8, 11, 12]. Some of the studies that used ClustalW were the papers of [2, 7]. However, progressive based MSA tools such as MUSCLE [5], ClustalW [8] and T-COFFEE [11], are likely to get trapped into local minima which is caused by its greedy nature [8, 12]. Studies by [5, 8, 12] stated that errors at early stages in the alignment cannot be corrected later which can propagate to the final alignment and may increase the likelihood of misalignments.

Iterative algorithms, on the other hand, can overcome the mentioned limitations of progressive algorithms but they have their own drawbacks. These algorithms include genetic algorithm-based methodologies like MOSAStrE [6], MSA-GA [9] and SAGA [12]. But the mentioned algorithms are much slower and stochastic in approach in where results may vary between runs unlike progressive algorithms which are fast and deterministic, in the sense that they always provide the same result [9]. Another drawback is the dependency of the final answer on the quality of the seed solution [6]. Despite the mentioned drawbacks, iterative algorithms have an important advantage over progressive methods since they are independent of the objective function in which any objective function can be optimized without modifications to the alignment routine [6, 12]. Although many MSA methodologies have emerged, the choice of the most suitable aligner and efficient evaluation method to measure alignment accuracy are still a problem [6].

For the scoring methods, substitution matrices are used. Substitution matrices provide a measure of the probability of a substitution (probability that a residue will be substituted by another residue due to mutation) or conservation (residue will not mutate) occuring [9]. For protein sequences, the most commonly used matrices are PAM [3] and BLOSUM [10] which only consider nucleotide or amino acid information to evaluate every aligned pair of residues. However, the mentioned matrices can lead to inaccurate alignments when the number of sequences increases or sequences are longer and more distant sequences are included [6]. Therefore, current scores are using additional information such as homologies or protein structures to complement alignment evaluations. Such scores include the STRIKE score [13] which was included in this study to measure alignment accuracy.

This study aims to find the appropriate parameters for MOSAStrE proposed in [6] since the authors have only used a single combination of crossover and mutation probabilities through out their simulations. MOSAStrE is a genetic algorithm based aligner implemented through the NSGA-II algorithm [1] and optimizes three objectives: percentage of non-gaps, percentage of totally conserved columns and the STRIKE score.

2 Materials and Methods

Multiobjective Optimizer for Sequence Alignments based on Structural Evaluations (MO-SAStrE) is a genetic algorithm based aligner implemented through the NSGA-II approach. In this study, input sequences were obtained from the BAliBASE dataset (v3.0). Initially, the alignment which we wish to optimize is fed to six different alignment tools. The resulting six alignments from these tools will be used to generate the initial parent population N_p. First, we include the six alignments to the initial population and the remaining N_p-6 individuals are generated by applying crossover. The new individuals (offspring) are built using the crossover and mutation operators according to the chosen crossover and mutation probabilities P_c and P_m respectively. The existing population and the newly built individuals are combined which are then referred to us the extended population. The best individuals, which will compose the next generation, are then selected from the extended population using the NSGA-II algorithm and will undergo the same process until the termination conditions are satisfied. The flow of the whole procedure is shown in Figure 1.

MO-SAStrE used three different objective functions to evaluate each alignment: STRIKE score, percentage of totally conserved (TC) columns and percentage of non-gaps. According to the STRIKE evaluation, contacts between amino acids in a sequence with known structure are estimated. For the remain-

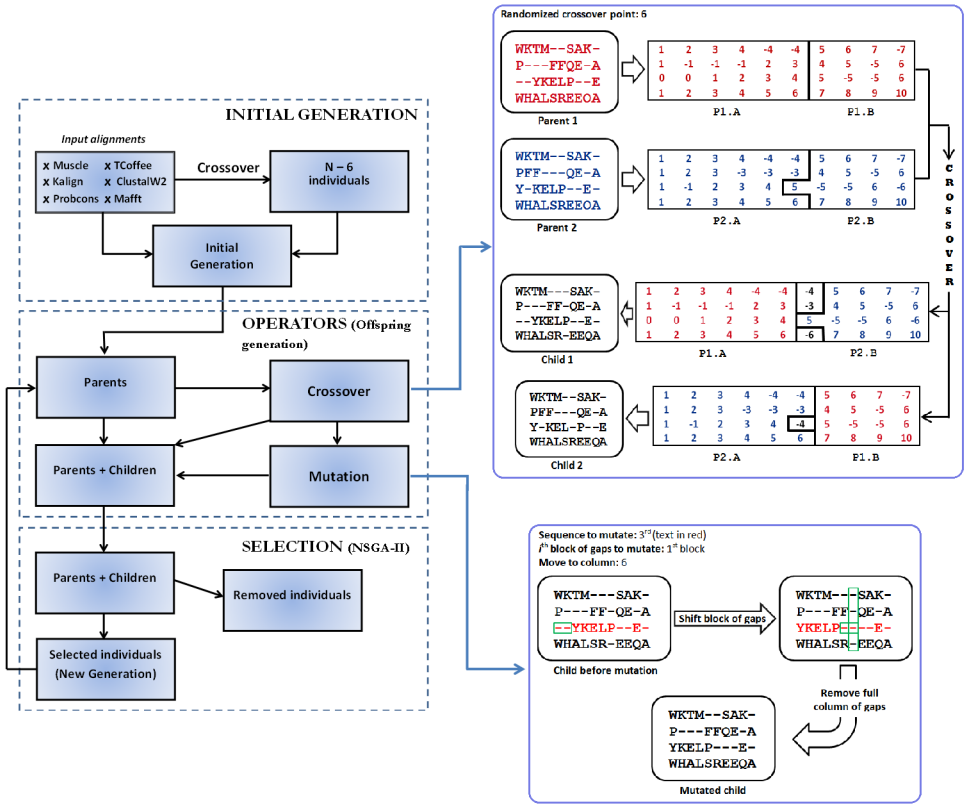

Figure 1: Flowchart for the five different test function values

ing sequences, the pairs of amino acids aligned in the same positions as the previously estimated contacts are retrieved and are scored using the STRIKE matrix provided by the STRIKE authors [13]. A high STRIKE score implies a higher structural similarity between the sequences. For the percentage of TC columns, it calculates the number of columns that are completely aligned with exactly the same amino acids. Higher TC values means the sequences have closer evolutionary relationships. The percentage of non-gaps, on the other hand, is the percentage of amino acid with respect to the number of gaps in the multiple sequence alignment. The NGP measures how compact and realistic an alignment is. Therefore, MO-SAStrE optimizes alignments by maximizing structural similarities between the sequences to be aligned and the number of totally conserved columns while reducing the number of gaps in the alignment.

The MO-SAStrE process terminates when either one of the three terminating conditions is satisfied. The first condition is the maximum number of generations allowed which is set to five hundred. Second is when the best STRIKE score do not change for twenty consecutive generations. The last condition is when the STRIKE scores of all individuals in the population are

equal and does not change for five consecutive generations.

3 Results and Discussion

Multiple runs were done to attain results for this parameter identification problem. Two sets of simulations were done on five datasets obtained from BAliBASE in order to find the appropriate parameters for MO-SAStrE. The first set maximizes three objective functions: STRIKE score, percentage of TC columns and percentage of non-gaps. While the second set of simulations only maximizes the STRIKE score to see if the structural similarities between the sequences aligned can be improved further. The five datasets, BB11001, BB11008, BB11009, BB11013 and BB11025, were taken from Ref. 1 v.1 of BAliBASE which are harder to align since sequences in each dataset are less similar.

The population was set to one hundred. The crossover-mutation probabilities, $P_c - P_m$ combination, were also varied according to the following combinations: 0.5 - 0.5, 0.6 - 0.4, 0.7 - 0.3, 0.8 - 0.2 and 0.9 - 0.1. Ten runs per combination of P_c and P_m were done on each of the five test datasets using the single-objective and multi-objective simulations yielding a total of five hundred (500) runs.

Table 1: Performance, runtime and convergence averages of different $P_c - P_m$ parameter combinations

Performance Criteria	Crossover rate-Mutation rate				
	0.5-0.5	0.6-0.4	0.7-0.3	0.8-0.2	0.9-0.1
STRIKE score	3.459	3.359	3.311	3.327	3.108
NGP (%)	59.790	59.361	58.565	58.825	59.488
TCP (%)	1.377	1.233	1.228	1.230	1.341
Runtime (mins.)	83.068	75.068	70.145	75.419	64.784
No. of generations	165.24	146.48	138.28	136.98	100.68

Since the multi-objective procedure returns the subset of non-dominated alignments which are equally good, it is not possible to decide which one is more accurate according to the three objectives [6]. In our results, we considered the alignment with the highest STRIKE score, which gives an alignment of more quality according to the sequence structures, as the best alignment. If in case the user selects the alignment with the highest non-gaps percentage (NGP), a more compact and realistic alignment can be obtained while an alignment with a high percentage of totally conserved columns (TCP) provides a better quality in terms of the evolutionary homologies among sequences.
Results of the simulations obtained from the multi-objective algorithm have

shown that the 0.5-0.5 generated the highest average objective function value among the five combinations. It had an average STRIKE score of 3.459, average non-gaps percentage of 59.790% and average percentage of totally conserved columns of 1.377% as shown in Table 1. Based on our simulations, the 0.8 crossover and 0.2 mutation rates used by MOSAStrE [6] was improved by 4% in the 0.5-0.5 combination. Similary, for the NGP scheme, the 0.5-0.5 combination has improved the score by 1.64%. For the TCP scheme, the 0.5-0.5 combination has improved the score by 11.95%. This implies that the alignments obtained from the said combination have the best qualities in terms of sequence structures and evolutionary homologies and are the most realistic ones among the alignments obtained using the other combinations of crossover and mutation probabilities. Therefore, the 0.5-0.5 combination which was ranked first in all the objective values is a more appropriate combination based on the simulations done on the five test datasets.

Moreover, results had shown that the simulations maximizing the three objectives all at once have higher objective values but are more than twice slower than the simulations maximizing only the STRIKE score. This only implies that the value of the STRIKE score cannot be improved further. Therefore, alignments obtained from the multiobjective algorithm are better not only in terms of the quality of sequence structures but are also far superior when it comes to achieving realistic alignments and alignments of greater evolutionary homologies though the algorithm is much slower.

4 Conclusion

MOSAStrE, a multiobjective genetic algorithm based sequence optimizer implemented through the NSGA-II algorithm, has been shown to be efficient in achieving optimized multiple sequence alignments based on the 0.8-0.2 combination of crossover and mutation probabilities used by the MOSAStrE authors. In this study, results obtained from the simulations have shown that the combination 0.5 crossover and 0.5 mutation rates gave the most realistic alignments and with the best qualities in terms of sequence structures and evolutionary homologies. Thus, it can be concluded that the said combination can be the most appropriate parameter to use rather than the 0.8 crossover and 0.2 mutation probabilities used by the MOSAStrE authors in expense of a higher computing time. But since only five datasets were used for the simulations and were only taken from a single Reference in BAliBASE, more simulations are needed to be done in order to show the efficiency of the 0.5-0.5 in other datasets.

Acknowledgement. The authors would like to thank the University of the Philippines for the support.

References

[1] S. Agarwal, K. Deb, T. Meyarivan and A. Pratap, A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II, *IEEE Transactions on Evolutionary Computation,* **6** (2002), 182-197. http://dx.doi.org/10.1109/4235.996017

[2] R. Bawane, I. Gontia, M. Kadam-Bedekar, S. Kumar, L.P.S. Rajput and K. Tantwai and S. Tiwari, Molecular Analysis of Phytase Gene Cloned from Bacillus subtilis, *Advanced Studies in Biology,* **3** (2011), 103-110.

[3] M.O. Dayhoff, B. C. Orcutt and R. M. Schwartz, A model of evolutionary change in proteins, *Atlas of Protein Sequence and Structure,* **5** (1978), 345-352.

[4] T. D. Duong, L. H. Ham, T. D. Khanh, N. T. Khoa, and K. H. Trung, Molecular Phylogeny of the Endangered Vietnamese Paphiopedilum Species Based on the Internal Transcribed Spacer of the Nuclear Ribosomal DNA, *Advanced Studies in Biology,* **5** (2013), 337-346. http://dx.doi.org/10.12988/asb.2013.3315

[5] R. C. Edgar, Muscle: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Research,* **32** (2004), 1792-1797. http://dx.doi.org/10.1093/nar/gkh340

[6] J.P. Florido, F.M. Ortuño, H. Pomares, F. Rojas, I. Rojas, J.M. Urquiza and O. Valenzuela, Optimizing multiple sequence alignments using a genetic algorithm based on three objectives: structural information, non-gaps percentage and totally conserved columns, *BMC Bioinformatics,* **29** (2013), 2112-2121. http://dx.doi.org/10.1093/bioinformatics/btt360

[7] P. Garcia-Fraile, P. F. Mateos and R. Rivas, Phylogenetic Diversity of Fast-Growing Basteria Isolated from Superficial Water of Lake Martel, a Saline Subterranean Lake in Mallorca Island (Spain) Formed by Filtration from the Mediterranean Sea through Underground Rocks, *Advance Studies in Biology,* **1** (2009), 333-344.

[8] T.J Gibson, D.G. Higgins and J.D. Thompson, 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** (1994), 4673-4680. http://dx.doi.org/10.1093/nar/22.22.4673

[9] C. Gondro and B.P. Kinghorn, A simple genetic algorithm for multiple sequence alignment,*Genetics and Molecular Research,* **6** (2007), 964-982.

[10] J.G. Henikoff and S. Henikoff, Amino-acid substitution matrices from protein blocks, *Proc. Natl. Acad. Sci. USA,* **89** (1992), 10915-10919. http://dx.doi.org/10.1073/pnas.89.22.10915

[11] J. Heringa, D. G. Higgins and C. Notredame, T-coffee: A novel method for fast and accurate multiple sequence alignment, *J. Mol. Biol.,* **302** (2000), 205-217. http://dx.doi.org/10.1006/jmbi.2000.4042

[12] D. Higgins and C. Notredame, SAGA: sequence alignment by genetic algorithm, *Nucleic Acids Research,* **24** (1996), 1515-1524. http://dx.doi.org/10.1093/nar/24.8.1515

[13] C. Kemena, J. Kleinjung, C. Notredame and J.F. Taly, STRIKE: evaluation of protein MSAs using a single 3D structure, *BMC Bioinformatics,* **27** (2011), 3385-3391. http://dx.doi.org/10.1093/bioinformatics/btr587

Received: December 22, 2015; Published: February 2, 2016