Computational Biology and Chemistry 75 (2018) 39–44

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

An approach for COFFEE objective function to global DNA multiple sequence alignment

Anderson Rici Amorim, Leandro Alves Neves, Carlos Roberto Valêncio, Guilherme Freire Roberto, Geraldo Francisco Donegá Zafalon*

Department of Computer Science and Statistics, São Paulo State University, Rua Cristovão Colombo 2265, São José do Rio Preto, São Paulo, Brazil

ARTICLE INFO

Article history:
Received 14 September 2016
Received in revised form 29 March 2018
Accepted 20 April 2018
Available online 25 April 2018

MSC:
00-01
99-00

Keywords:
Multiple sequence alignment
Genetic Algorithm
Optimization

ABSTRACT

Multiple sequence alignment (MSA) is one of the most important tasks in bioinformatics and it can be used to prediction of structures or functions of unknown proteins and to phylogenetic tree reconstruction. There are many heuristics to perform multiple sequence alignment, as Progressive Alignment, Ant Colony, Genetic Algorithms, among others. Along the years, some tools were proposed to perform MSA and MSA-GA is one of them. The MSA-GA is a tool based on Genetic Algorithm to perform multiple sequence alignment and its results are generally better than other well-known tools in bioinformatics, as Clustal W. The COFFEE objective function was implemented in the MSA-GA in order to allow it to produce better alignments to less similar sequence sets of proteins. Nonetheless, the COFFEE objective function is not suited do perform multiple sequence alignment of nucleotides. Thus, we have modified the COFFEE objective function, previously implemented in the MSA-GA, to allow it to obtain better results also to sequences of nucleotides. Our results have shown that our approach has achieved better results in all cases when compared with standard COFFEE and most of cases when compared with WSP for all test cases from BAliBase and BRAliBase. Moreover, our results are more reliable because their standard deviations have less variation.

1. Introduction

Due the rise of genomic data currently available, the multiple sequence alignment has been considered one of the most important tasks of bioinformatics. It plays an important role in sequence analysis, as function prediction of unknown protein structures (Pei and Grishin, 2014; Li et al., 2015), viral genome decoding (Greive et al., 2016), diseases studies (Jordan et al., 2015), among others.

The better alignment of given sequences can be deterministically obtained by dynamic programming algorithms, as Needleman–Wunsch (Needleman and Wunsch, 1970) or Smith–Waterman (Smith and Waterman, 1981). The first one is a well-known algorithm to global sequence alignments and the second one is used to execute local sequence alignments. However, these algorithms were ideally developed to align pair of sequences and, due their computational costs, it is infeasible produce an alignment for three or more sequences (Wang and Jiang, 1994).

Thus, to smooth the high computational cost, the multiple sequence alignment algorithms were proposed. This programs can be based on many heuristics, as Ant Colony (Lee et al., 2008), Simulated Annealing (Yao et al., 2015), Progressive Alignment (Sievers and Higgins, 2014), Genetic Algorithms (GA) (Zhu et al., 2016), among others.

Concerning GA, it can be used to solve multiple sequence alignment problems through an approach based on Evolution Theory (Yadav and Banka, 2016). In this approach, individuals of a population are exposed to mutation, recombination and selection to evolve a population of possible solutions whose biological significances are measured by an objective function (Notredame and Higgins, 1996).

The first tool developed to solve multiple sequence alignment problems using GA was SAGA (Sequence Alignment by Genetic Algorithm) (Notredame and Higgins, 1996). It has a complex pack of 22 operators, including mutation and recombination, which are selected by an automatic scheduler. However, some studies have shown that the complexity of SAGA is unnecessary and the automatic scheduler does not improve the final quality of the alignments when compared to an uniform selector of operators (Thomsen and Boomsma, 2004).

* Corresponding author.
E-mail address: geraldo.zafalon@unesp.br (G.F.D. Zafalon).

https://doi.org/10.1016/j.compbiolchem.2018.04.012
1476-9271/

Another possibility of tool that uses GA in multiple sequence alignment of amino acids and nucleotide sequences is MSA-GA (Gondro and Kinghorn, 2007). This tool is more efficient to solve global multiple sequence alignment problems because of its simpler computational approach. In addition, it produces better results when compared with other well-known tools in bioinformatics, as Clustal W (Thompson et al., 1994).

However, the default objective function used by MSA-GA is Weighted Sum-of-Pairs (WSP). This objective function, when evaluates less similar sequence sets, generally produces noisy alignments, which is, sometimes, undesirable. Thus, Amorim et al. (2015) implemented the COFFEE (Consistency based Objective Function For alignmEnt Evaluation) objective function (Notredame et al., 1998) in MSA-GA, which has allowed the tool to produce, in general, better results than the original approach.

Nonetheless, the COFFEE objective function was developed to evaluate protein sequences (Notredame et al., 1998) and its implementation to evaluate nucleotide sequences is inadequate. Thus, Wang and Lefkowitz (2005) modified the COFFEE function to apply it to local sequence alignments, specifically in inconsistent regions. However, this modification has not been extended to global multiple sequence alignment of nucleotides.

Thus, to expand the improvements obtained by Amorim et al. (2015), we have modified the COFFEE objective function to allow its implementation on tools to global multiple sequence alignment, as MSA-GA. In the results we have obtained, we are able to notice the improvements developed can produce better results to global multiple sequence alignments of nucleotides using COFFEE as evaluation scheme.

2. Materials and methods

2.1. MSA-GA and COFFEE objective function

The MSA-GA is a tool to perform multiple sequence alignments of amino acids or nucleotides using a simple GA and the WSP function to evaluate the quality of the obtained results (Gondro and Kinghorn, 2007). In bioinformatics, the MSA-GA stands out because it can produce better results when compared with other widely used tools, as Clustal W. However, due to the WSP nature, MSA-GA generally cannot produce results with good biological significance when it aligns less similar sequence sets (Amorim et al., 2015).

Thus, as Notredame et al. (1998) identified some noisy problems in the final results of well-done objective functions, where WSP is included, they developed the COFFEE, which is based on consistency with a pairwise library, smoothing the disadvantages of those functions in order to produce better results. The

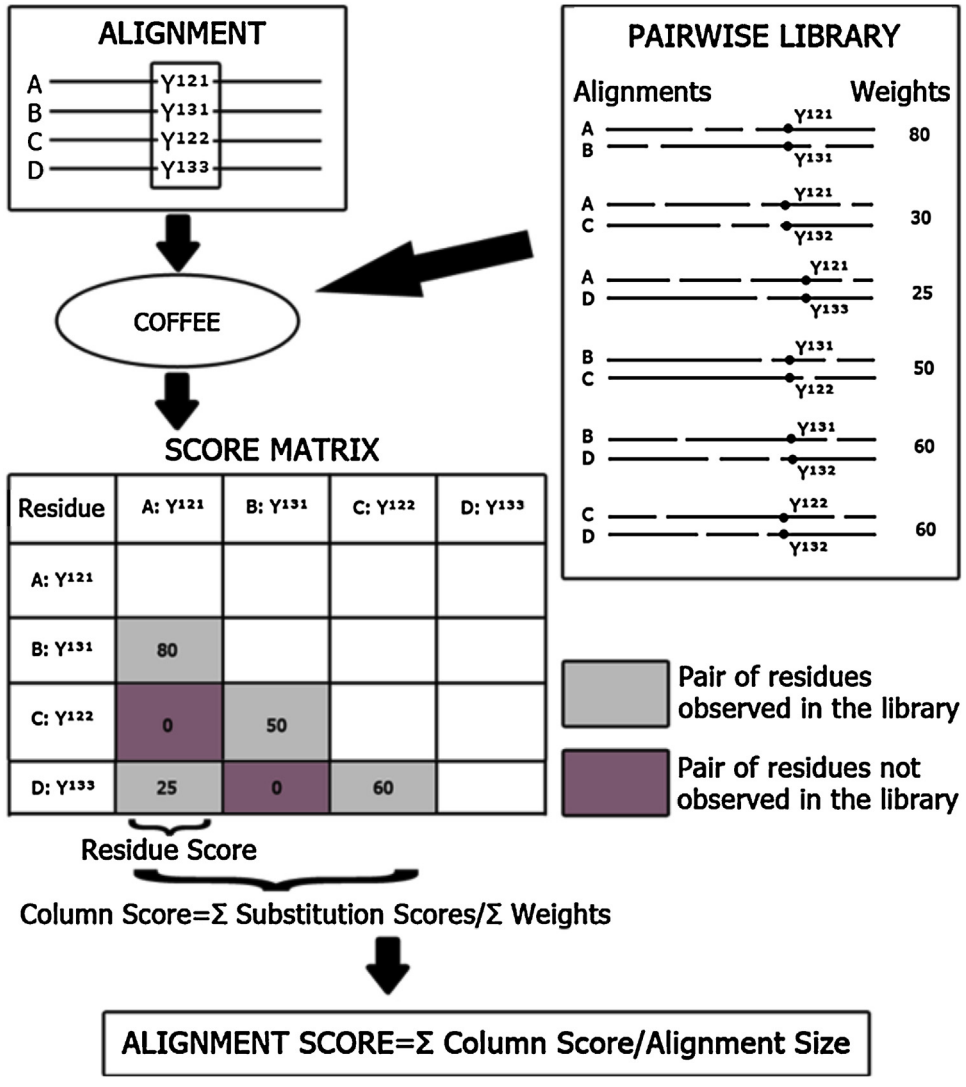

Fig. 1. COFFEE objective function scheme.

COFFEE objective function is formalized by Eq. (1), where N is the number of sequences, $S_1, \ldots, S_N$, in a multiple sequence alignment, $LEN(A_{(i,j)})$ is the alignment length, $SCORE(A_{(i,j)})$ is the number of pair of residues shared between the library and $A_{(i,j)}$ and, finally, $W_{(i,j)}$ is the weight associated to the correspondent pairwise alignment.

$$\text{COFFEE score} = \frac{\left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} W_{i,j} * SCORE(A_{i,j})\right]}{\left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} W_{i,j} * LEN(A_{i,j})\right]} \quad (1)$$

In order to make MSA-GA more robust, Amorim et al. (2015) have implemented the COFFEE objective function in this tool to evaluate the individuals of the GA. This modification allowed the MSA-GA to produce alignments of less similar sequence sets with good biological significance.

Nonetheless, COFFEE is not suited to evaluate multiple sequence alignment of nucleotides, producing, in these cases, poor quality results (Notredame et al., 1998). Thus, we have modified the routine implemented by Amorim et al. (2015) in the MSA-GA tool, in order to allow that this method directs the tool's genetic algorithm to improve the alignment of nucleotides in terms of biological significance.

2.2. Improvements in the COFFEE function to global multiple sequence alignment of nucleotides

Analyzing the modifications made by Amorim et al. (2015), it can be noticed that the COFFEE function was implemented using two main routines called *MountMatrix* and *SearchInSeq*. The first one declares and fills the score matrix of each alignment and the second one searches for correspondences between the analyzed pair of residues and the pairwise library. In this approach, Amorim et al. (2015) have obtained improvements in the sensibility of the MSA-GA tool when it aligns less similar sequence sets. However, to extend these improvements to the nucleotide alignment module, it was necessary an adaptation in the COFFEE objective function.

Thus, we have modified the *SearchInSeq* function in order to allow COFFEE to produce better alignments of nucleotide sequences. Originally, the COFFEE function scores pairs of residues if they are found in any position of the corresponding alignment at the pairwise library, as can be seen in Fig. Figure 1.

In the present work, we have modified the COFFEE objective function, which is presented in Fig. 2, in order to restrict the search for an optimal alignment. Thus, we made changes in the *SearchInSeq* function, in order to score just the pair of residues

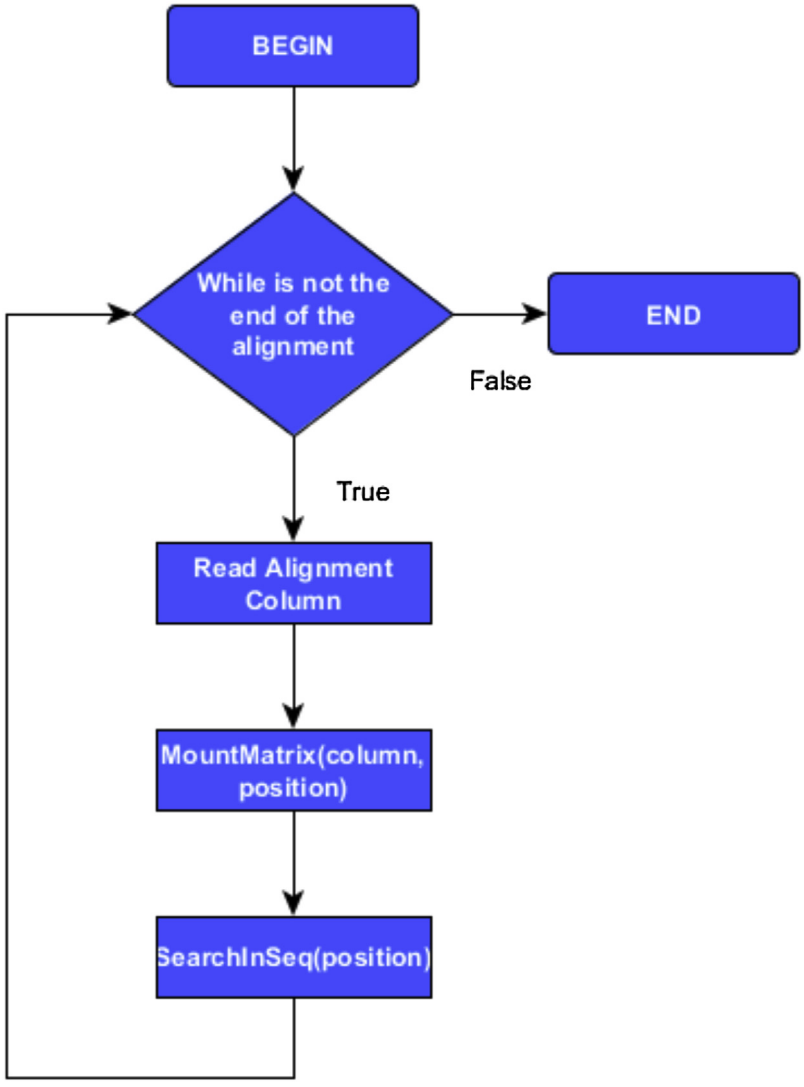

Fig. 2. COFFEE objective function flowchart. The current position of the alignment column being analyzed is a parameter to *MountMatrix* and *SearchInSeq*.

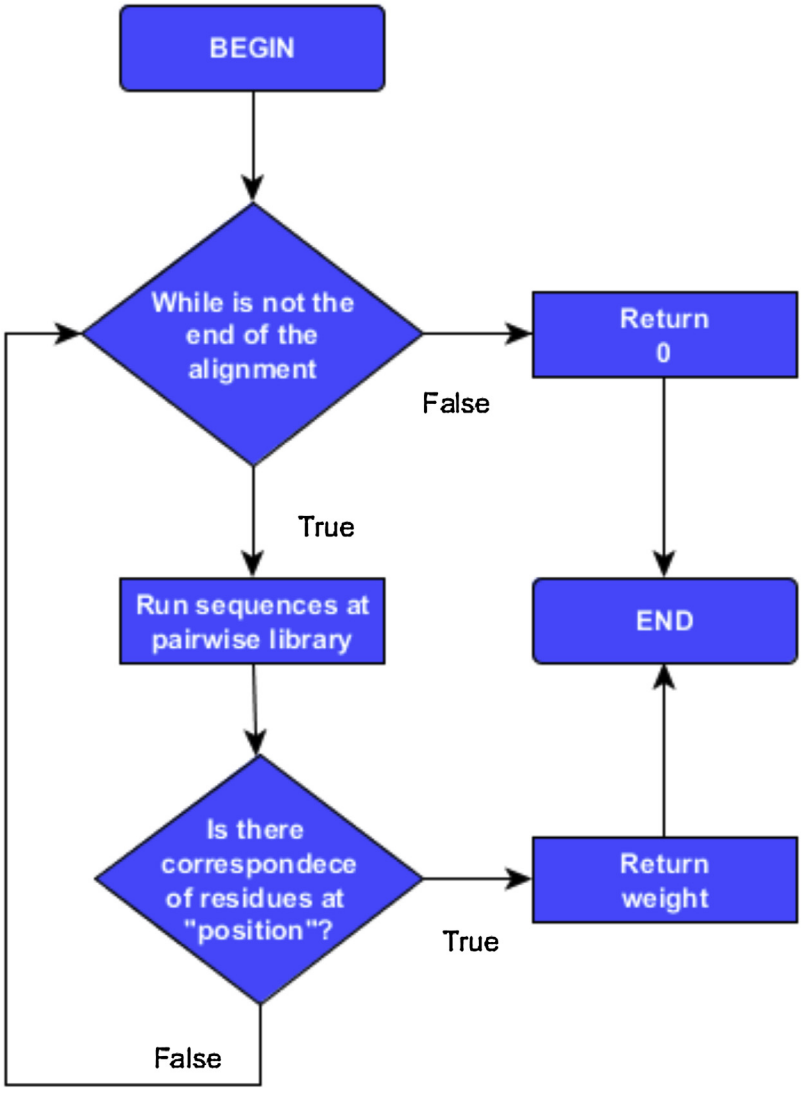

Fig. 3. *SearchInSeq(position)* function flowchart.

that have correspondence at the pairwise library exactly at the same position of the column that are being analyzed on the multiple sequence alignment, as can be seen in the flowchart presented in Fig. 3. To improve the comprehension of the flowchart presented in Fig. 3, specially to describe the correspondence of the position in the MSA and in the pairwise alignment in the library, an illustration is presented in Fig. 4

The original concept of COFFEE was to find the correspondence of a base pair (residue) in any position of the pairwise alignment of the library. The aim of COFFEE objective function is

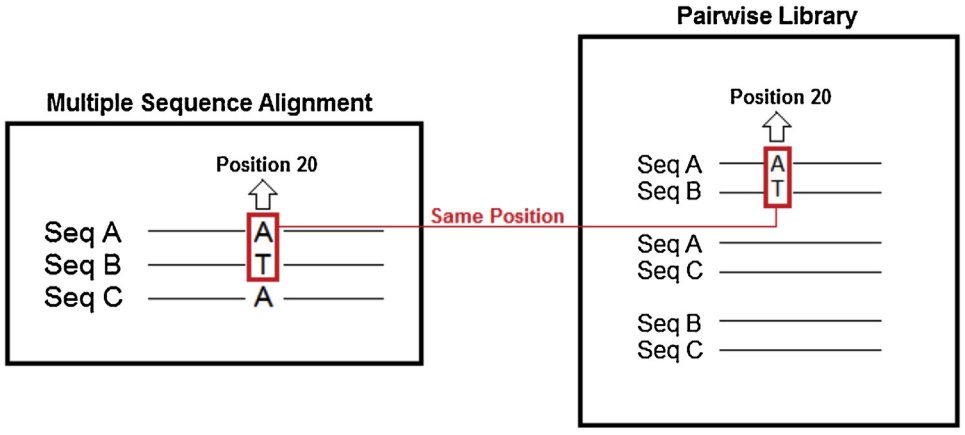

Fig. 4. Illustration of the position lock.

Table 1
Average scores of MSA-GA with the standard COFFEE, modified COFFEE and WSP for all test cases of BAliBase.

Subset	COFFEE-S	COFFEE-M	WSP
Ref. 1	0.046	**0.405**	0.369
Ref. 2	0.058	**0.344**	0.311
Ref. 3	0.052	**0.257**	0.232
Ref. 4	0.012	**0.065**	0.057
Ref. 5	0.064	**0.263**	0.196

Table 2
Average scores of MSA-GA with the standard COFFEE, modified COFFEE and WSP for all test cases of BRAliBase.

Subset	COFFEE-S	COFFEE-M	WSP
G2Intron	0.472	**0.702**	0.603
rRNA	0.634	**0.835**	0.795
tRNA	0.683	**0.772**	0.759
U5	0.452	**0.628**	0.603

to ensure that multiple alignment reaches the maximum number of correspondences when compared with the pairwise alignments in the library. The proposed modification also aims to find the correspondences, however, ensuring these correspondences occur in the same position both in the multiple alignment and in the pairwise alignment in the library. The position lock is a way to improve the consistency of results, because the location of occurrence becomes restricted. Analyzing from the mathematical point of view, the original COFFEE does not work properly with the new approach, because as it was developed to work with a bigger set of elements (amino acids) than the new proposed approach (nucleotides) there is a reduction in the probability of finding a correspondence in the same location both for multiple alignment and for pairwise alignment in the library. Thus, this probability is considerably improved when the set of elements is reduced (nucleotides only), because the chance of finding the correspondence in the same location increases with lower range of elements. Therefore, the obtained results of modified COFFEE are better than original COFFEE when nucleotide sequences are used.

As the main idea of COFFEE function is to find a very similar solution to the exact alignments in the pairwise library, our adaptations were essentials to allow MSA-GA producing global multiple sequence alignment with good biological significance, extending the improvements obtained by Amorim et al. (2015) to the alignment of nucleotides module as well.

3. Results and discussion

3.1. Benchmark and test platform

In order to evaluate the quality of the obtained solutions with the modifications presented here, we have used test cases of the reverse transcription of proteins to DNA from BAliBase (Thompson et al., 2005; Carroll et al., 2007), which are freely available on the web.[1] The BAliBase is a well suited benchmark to evaluate the quality of the results obtained by multiple sequence alignment tools, because it has different test cases classified through categories with different similarities, which are essentials to a satisfactory evaluation of the results produced by MSA-GA with the modified COFFEE objective function.

Moreover, we have used in the tests the BRAliBase benchmark (Gardner et al., 2005) to make the quality evaluation more robust. This benchmark provides sequence sets of nucleotides and their respective structural alignments, which allows comparing the obtained results with the correct reference alignment. The BRAliBASE is divided into four groups: G2Intron, rRNA, tRNA and U5, and each of them contains different test cases with different characteristics.

To evaluate the biological significance of the produced alignments, we have used the tool *qscore*, which is available along with BAliBase. Among the several scores offered by *qscore* we have used PREFAB Q one. This score compares the obtained alignment with a reference alignment and returns a score between one and zero, where zero is the worst alignment and one is the best.

All tests were executed using a Dell Vostro computer with Windows 8.1 Pro 64 bits, Intel Core i5-3470S CPU@2.90GHz processor and 6GB of RAM memory. The parameters used in the MSA-GA were the same used by Gondro and Kinghorn (2007).

3.2. Quality tests

Concerning the quality tests in this work, we have executed all test cases of the References 1, 2, 3, 4 and 5 from BAliBase and all test cases of the groups G2Intron, rRNA, tRNA and U5 from BRAliBase, with different characteristics, as sequence length and similarity level. All of these tests that we have performed for BAliBase[2] and BRAliBase[3] are available to download from everywhere.

Due to the fact of the stochastic approach of the GA, it generally produces different alignments for the same sequence set. Then, we have executed each test case five times, in order to ensure that the obtained results were statistically correct. Therefore, the score considered in the present work for each test case is the average of the scores obtained through all executions.

In Table Table 1 are presented the average scores obtained by the execution of all test cases from each Reference of BAliBase. The tests were performed for standard COFFEE (COFFEE-S), modified COFFEE (COFFEE-M) and WSP approaches.

As can be noticed in Table Table 1, all the obtained results of COFFEE-M are better than COFFEE-S and WSP. For all results, when we analyze the improvement of COFFEE-M in relation to the COFFEE-S, we obtained an average of 584%, and 15.8% for COFFEE-M in relation to WSP. To obtain these averages we have calculated the improvement for each Reference (Ref. 1, Ref. 2, Ref. 3, Ref. 4 and Ref. 5), comparing COFFEE-M with COFFEE-S and COFFEE-M with WSP. After that, with the improvements of all References, we have calculated the average improvement of them. In this case, the average standard deviation of the standard COFFEE was 0.024, of the modified COFFEE was 0.020, while to WSP was 0.040.

In Table Table 2 are presented the average scores obtained by the execution of all test cases from each group of BRAliBase. The tests were performed for standard COFFEE (COFFEE-S), modified COFFEE (COFFEE-M) and WSP approaches.

As can be noticed in Table Table 2, all the obtained results of COFFEE-M are better than COFFEE-S and WSP. For all results, when we analyze the improvement of COFFEE-M in relation to the COFFEE-S, we obtained an average of 33.1%, and 6.8% for COFFEE-M in relation to WSP. In this case, the average standard deviation of the standard COFFEE was 0.115, of the modified COFFEE was 0.089, while to WSP was 0.101.

[1] http://www.drive5.com/bench.

[2] http://www.ibilce.unesp.br/gcc/balibase-results-final.xlsx.

[3] http://www.ibilce.unesp.br/gcc/bralibase-results-final.xlsx.

4. Conclusion

The MSA-GA is a widespread multiple sequence alignment tool because its simple GA scheme is capable of produce better results when compared with other well-known tools, as Clustal W (Gondro and Kinghorn, 2007). However, the standard objective function of MSA-GA is the WSP, which generally cannot produce quality alignments when evaluating sets with less similar sequences.

Thus, Amorim et al. (2015) implemented the COFFEE objective function in the MSA-GA tool to smooth the disadvantage previously referred. However, the COFFEE function was ideally developed to evaluate multiple sequence alignment of amino acids. So, in order to extend the obtained improvements to the DNA alignment module, we modified the COFFEE to produce also good results to alignments of nucleotides. Basically, we have improved the function to score only the pair of residues consistent with the pairwise library, at the exact position of the analyzed multiple sequence alignment's column.

Thus, considering the reverse transcription of BAliBase in all its data sets, our modified COFFEE was able to produce better results than standard COFFEE in 100% of the times and in 71.3% of the times when compared with WSP. Concerning the average quality improvement, the modified COFFEE achieved 584% in relation to standard COFFEE and 15.8% in relation to WSP, in terms of biological significance. Moreover, when we analyze the results from the execution of all BRAliBase data sets, we can see that our modified COFFEE was able to produce better results than standard COFFEE in 100% of the times and in 87.6% of the times when compared with WSP. Concerning the average quality improvement, the modified COFFEE achieved 33.1% in relation to standard COFFEE and 6.8% in relation to WSP, in terms of biological significance. Finally, the new approach with the modified COFFEE presented less variation in its results when compared with standard COFFEE and WSP, which ensures more consistency in the results and improves the biological analysis.

Acknowledgment

The authors would like to thank all of our collaborators and our institutions for the support to the development of the present work and São Paulo Research Foundation (FAPESP) under grant number 13/08289-0.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.compbiolchem.2018.04.012.

References

Amorim, A.R., Zafalon, G.F.D., Neves, L.A., Pinto, A., Valêncio, C.R., Machado, J.M., 2015. Improvements in the sensibility of MSA-GA tool using coffee objective function. J. Phys.: Conf. Ser. 574 (1), 012104.
Carroll, H., Beckstead, W., O'connor, T., Ebbert, M., Clement, M., Snell, Q., McClellan, D., 2007. DNA reference alignment benchmarks based on tertiary structure of encoded proteins. Bioinformatics 23 (19), 2648–2649.
Gardner, P.P., Wilm, A., Washietl, S., 2005. A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res. 33 (8), 2433–2439.
Gondro, C., Kinghorn, B., 2007. A simple genetic algorithm for multiple sequence alignment. Genet. Mol. Res. 6 (4), 964–982.
Greive, S.J., Fung, H.K., Chechik, M., Jenkins, H.T., Weitzel, S.E., Aguiar, P.M., Brentnall, A.S., Glousieau, M., Gladyshev, G.V., Potts, J.R., et al., 2016. DNA recognition for virus assembly through multiple sequence-independent interactions with a helix-turn-helix motif. Nucleic Acids Res. 44 (2), 776–789.
Jordan, D.M., Frangakis, S.G., Golzio, C., Cassa, C.A., Kurtzberg, J., Davis, E.E., Sunyaev, S.R., Katsanis, N., et al., 2015. Identification of cis-suppression of human disease mutations by comparative genomics. Nature 524 (7564), 225–229.
Lee, Z.-J., Su, S.-F., Chuang, C.-C., Liu, K.-H., 2008. Genetic algorithm with ant colony optimization (GA-ACO) for multiple sequence alignment. Appl. Soft Comput. 8 (1), 55–78.
Li, B., Chiong, R., Lin, M., 2015. A balance-evolution artificial bee colony algorithm for protein structure optimization based on a three-dimensional ab off-lattice model. Comput. Biol. Chem. 54, 1–12.
Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48 (3), 443–453.
Notredame, C., Higgins, D.G., 1996. Saga: sequence alignment by genetic algorithm. Nucleic Acids Res. 24 (8), 1515–1524.
Notredame, C., Holm, L., Higgins, D.G., 1998. Coffee: an objective function for multiple sequence alignments. Bioinformatics 14 (5), 407–422.
Pei, J., Grishin, N.V., 2014. Promals3d: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. Methods Mol Biol 263–271.
Sievers, F., Higgins, D.G., 2014. Clustal omega, accurate alignment of very large numbers of sequences. Methods Mol. Biol. 105–116.
Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147 (1), 195–197.
Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22 (22), 4673–4680.
Thompson, J.D., Koehl, P., Ripp, R., Poch, O., 2005. Balibase 3.0: latest developments of the multiple sequence alignment benchmark. Proteins: Struct. Funct. Bioinformatics 61 (1), 127–136.
Thomsen, R., Boomsma, W., 2004. Multiple sequence alignment using saga: investigating the effects of operator scheduling, population seeding, and crossover operators. Workshops on Applications of Evolutionary Computation 113–122 Springer.
Wang, L., Jiang, T., 1994. On the complexity of multiple sequence alignment. J. Comput. Biol. 1 (4), 337–348.
Wang, C., Lefkowitz, E.J., 2005. Genomic multiple sequence alignments: refinement using a genetic algorithm. BMC Bioinformatics 6 (1), 1.
Yadav, R.K., Banka, H., 2016. Genetic algorithm using guide tree in mutation operator for solving multiple sequence alignment. Advanced Computing and Systems for Security 145–157 Springer.
Yao, D., Jiang, M., You, X., Abulizi, A., Hou, R., 2015. An algorithm of multiple sequence alignment based on consensus sequence searched by simulated annealing and star alignment. 2015 International Symposium on Bioelectronics and Bioinformatics (ISBB) 3–6 IEEE.
Zhu, H., He, Z., Jia, Y., 2016. A novel approach to multiple sequence alignment using multiobjective evolutionary algorithm based on decomposition. IEEE J. Biomed. Health Informatics 20 (2), 717–727.