

## Improvements in the sensibility of MSA-GA tool using COFFEE objective function

A R Amorim<sup>1</sup>, G F D Zafalon<sup>1</sup>, L A Neves<sup>1</sup>, A R Pinto<sup>2</sup>, C R Valêncio<sup>1</sup>, J M Machado<sup>1</sup>

<sup>1</sup> Department of Computer Science and Statistics (DCCE), São Paulo State University (UNESP), São José do Rio Preto, Brazil.

<sup>2</sup> Department of Control Engineering and Automation, Federal University of Santa Catarina, Blumenau, Brazil.

E-mail: anderson.rici@sjrp.unesp.br

**Abstract.** The sequence alignment is one of the most important tasks in Bioinformatics, playing an important role in the sequences analysis. There are many strategies to perform sequence alignment, since those use deterministic algorithms, as dynamic programming, until those ones, which use heuristic algorithms, as Progressive, Ant Colony (ACO), Genetic Algorithms (GA), Simulated Annealing (SA), among others. In this work, we have implemented the objective function COFFEE in the MSA-GA tool, in substitution of Weighted Sum-of-Pairs (WSP), to improve the final results. In the tests, we were able to verify the approach using COFFEE function achieved better results in 81% of the lower similarity alignments when compared with WSP approach. Moreover, even in the tests with more similar sets, the approach using COFFEE was better in 43% of the times.

### 1. Introduction

Some studies performed in biology have brought many improvements, specially in human genetics in the last years. However, these studies produce a huge amount of data, which must be refined for a more accurate analysis and after to propose some inferences.

Nowadays, the use of computational tools is necessary to provide a good genomic analysis, which can be classified as Bioinformatics. Thus, Bioinformatics is a computer science branch to solve biological problems, mainly related to multiple sequence alignments (MSA) and pattern recognition [1].

In this work, we proposed the implementation of COFFEE objective function in the multiple sequence alignment tool MSA-GA. Thus, it is possible to reach alignments with more biological significance for some sequence sets, specially those ones with lower similarity.

This work is organized as follows: in the section 2 a brief review about sequence alignment and genetic algorithm is provided. In the section 3 are described the materials and methods with the special attention to the implementation of the objective function in the MSA-GA tool. Some analysis and results are presented in the section 4. Finally, in the section 5, the conclusions and future perspectives are presented.



## 2. Sequence Alignment and Genetic Algorithms

The exact score of a sequence alignment can be obtained using dynamic programming algorithms, as Needleman-Wunsch [2]. However, this type of algorithm has high computational costs, becoming the alignments with more than two sequences unfeasible. Thus, to reduce the computational complexity, the multiple sequence alignment algorithms were developed, which can be based in many heuristics as Progressive Alignment, Simulated Annealing, Genetic Algorithms, and others [3].

Genetic Algorithms might be used to solve MSA problems through the Evolutionary Theory, where the participants are submitted to processes of mutation, recombination and gene selection to evolve the candidate alignments, which are measured by an objective function. Generally, in this heuristics, the alignment module is independent of score function, therefore the objective function can be implemented without changes in the alignment routines.

## 3. Materials and Methods

The MSA-GA [4] is a tool for MSA, which uses genetic algorithm as its base. The choice of this tool to the development of this work was done due to its good results when compared with other MSA tools. Moreover, the MSA-GA has a good modularity which allows the implementation of new objective functions in a feasible way. The MSA-GA uses the Weighted Sum-of-Pairs (WSP) as the objective function. However, the WSP has some limitations where regions with low similarity can cause some distortions in the final alignments.

In this context, a new score method proposed by Notredame [5], named COFFEE (Consistency based Objective Function For alignmEnt Evaluation), was implemented in the MSA-GA tool. The choice of this objective function was performed due to its approach is based in consistency, which smooths the noise caused by regions with low similarity, resulting in alignments with more biological significance even in sequences with low similarity.

Basically, the COFFEE function needs two components: the reference pairwise alignment set, which is called library, and an objective function which analyzes the consistence between a MSA and the pairwise alignments. In this work, the Needleman-Wunsch algorithm was used to build the library. To reduce the noise caused by sequences with low similarity, each pairwise alignment has an weight, which is related to the number of matches in the aligned sequences belonged to the library. The consequence is that the final alignment privileges closer sequences instead of more distant ones.

Thus, the function score routine is based in the comparison between each pair of aligned residues with those belonged to the library, as presented in the Figure 1. Basically, for each column of multiple alignment, a residue matrix is declared. If the pair of residues is found in the pairwise alignment, the matrix cell is filled with the weight of the alignment.

The score of each column is reached by sum of matrix elements, divided by sum of the alignment weights in the library. Finally, the general score of consistency is equal to the sum of scores of columns, divided by the numbers of pairs of multiple alignment.

The COFFEE function is defined by Equation (1), where  $N$  is the number of sequences,  $LEN(A_{i,j})$  is the length of the alignment,  $SCORE(A_{i,j})$  is the number of pairs of aligned residues shared between  $A_{i,j}$  and the library, and  $W_{i,j}$  is the weight of pairwise alignment.

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

## 4. Tests and Results

To analyze the improvements of this work, we used the test case sets from BALiBase [6]. This benchmark offers many sequence sets divided into different reference categories, where there is the possibility of an accurate comparison of obtained MSA with reference alignments. This comparison is performed using the BALiScore tool, which gives a score of biological significance, where 0 is the worst and 1 is the best alignment.

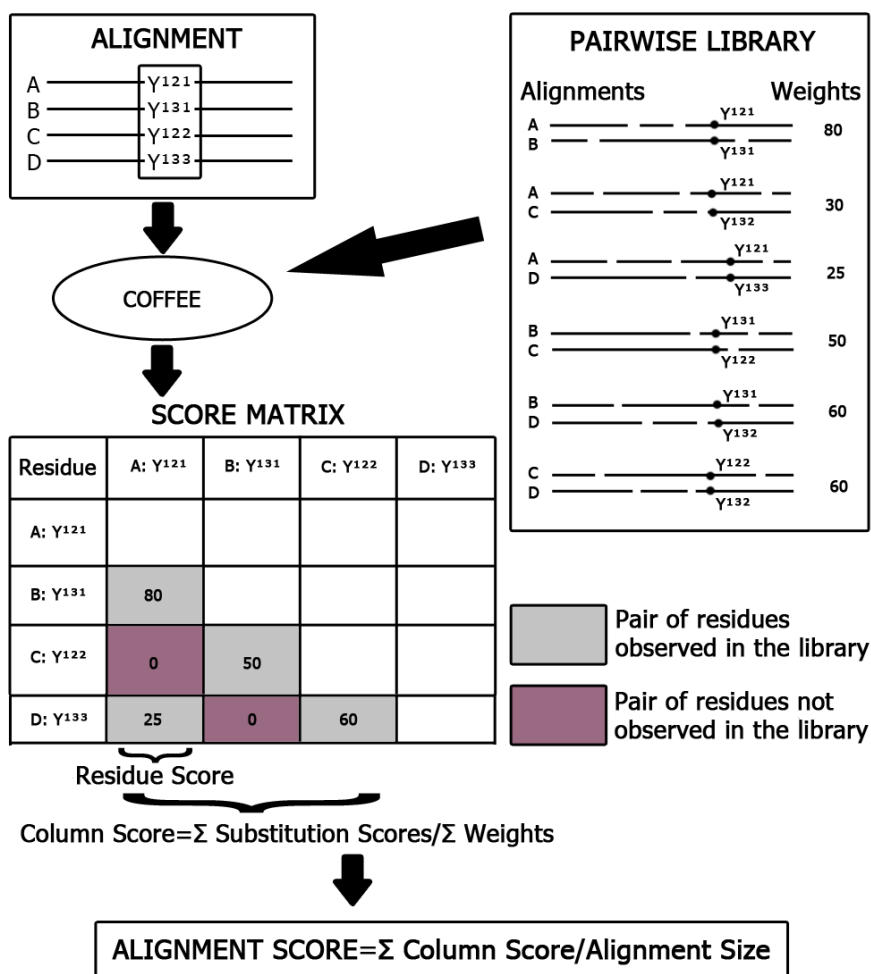


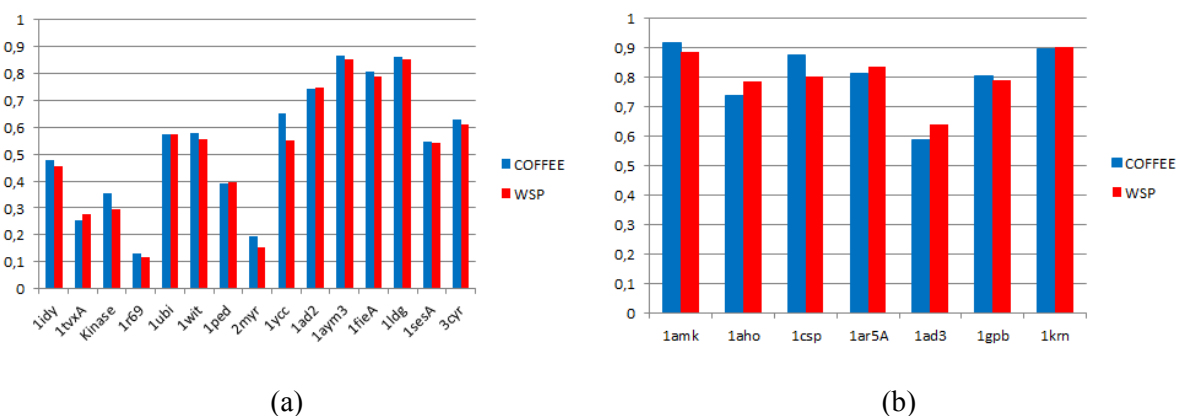
Figure 1. COFFEE function scheme.

In the tests, the cases of Reference 1 were selected, all with equidistant sequence sets with different conservation. The parameters used by genetic algorithm, in both approaches, were patterns of MSA-GA. Thus, in the Table 1 are shown the obtained scores of COFFEE comparing with WSP.

Table 1. COFFEE and WSP scores in the BALiScore.

Sequence Set	MSA-GA (COFFEE)	MSA-GA (WSP)	Sequence Set	MSA-GA (COFFEE)	MSA-GA (WSP)	Sequence Set	MSA-GA (COFFEE)	MSA-GA (WSP)
			<b>20% ~ 40% identity</b>					
			<b>&gt;35% identity</b>					
<b>&lt;25% identity</b>			lycc	<b>0,6520</b>	0,5488	lamk	<b>0,9176</b>	0,8874
lidy	<b>0,4788</b>	0,4538	lad2	0,7408	<b>0,7498</b>	laho	0,7386	<b>0,7848</b>
ltvxA	0,2532	<b>0,2748</b>	laym3	<b>0,8650</b>	0,8526	lcsp	<b>0,8754</b>	0,8028
Kinase	<b>0,3532</b>	0,2940	lfieA	<b>0,8058</b>	0,7902	lar5A	0,8142	<b>0,8376</b>
lr69	<b>0,1290</b>	0,1176	lldg	<b>0,8634</b>	0,8546	lad3	0,5872	<b>0,6394</b>
lubi	<b>0,5756</b>	0,5752	lsesA	<b>0,5476</b>	0,5416	lgbp	<b>0,8058</b>	0,7902
lwit	<b>0,5792</b>	0,5548	3cyr	<b>0,6302</b>	0,6088	lkrn	0,8964	<b>0,9052</b>
lped	0,3906	<b>0,3968</b>						
2myr	<b>0,1926</b>	0,1548						

Thus, it can be noticed the COFFEE function was able to improve the sensibility of the final alignment of MSA-GA, where in 81% of the cases, the new strategy proposed here has reached better results in sequence sets of low similarity (<25% identity, 20% ~ 40% identity), when compared with WSP as showed in Figure 2(a). Moreover, the COFFEE function has reached good results in sequence sets with high similarity (>35% identity), with better results in 43% of the cases, when compared with WSP, as can be seen in Figure 2(b).



**Figure 2.** BALiScore score: (a) sets with low similarity, (b) sets with high similarity.

## 5. Conclusion

The novelty proposed in this work is very important, because, through the implementation of COFFEE function into the MSA-GA, this tool was able to reach alignments with more biological significance, even for sequence sets with low similarity and also good results for those with high similarity. Moreover, some characteristics of genetic algorithm allow a parallel implementation of it, which will be conducted using multithreading strategy to reduce the execution time of the tool.

## 6. Acknowledgments

This work was financially supported by São Paulo Research Foundation (FAPESP), under grant 2013/08289-0.

## References

- [1] Edgar R C and Batzoglou S 2006 Multiple sequence alignment *Current Opinion in Structural Biology* **16** 368–373
- [2] Needleman S B and Wunsch C D 1970 A general method applicable to the search for similarities in the amino acid sequence of two proteins *Journal of molecular biology* **48** 443-453
- [3] Zafalon G F et al. 2013 Improvements in the score matrix calculation method using parallel score estimating algorithm *Journal of Biophysical Chemistry* **4** 47-51
- [4] Gondro C and Kinghorn B P 2007 A simple genetic algorithm for multiple sequence alignment *Genetics and Molecular Research* **6** 964-982
- [5] Notredame C, Holm L and Higgins D G 1998 COFFEE: an objective function for multiple sequence alignments *Bioinformatics* **14** 407-422
- [6] Thompson J D, Koehl P, Ripp R and Poch O 2005 BALiBASE 3.0: latest developments of the multiple sequence alignment benchmark *Proteins: Structure, Function, and Bioinformatics* **61** 127-136