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

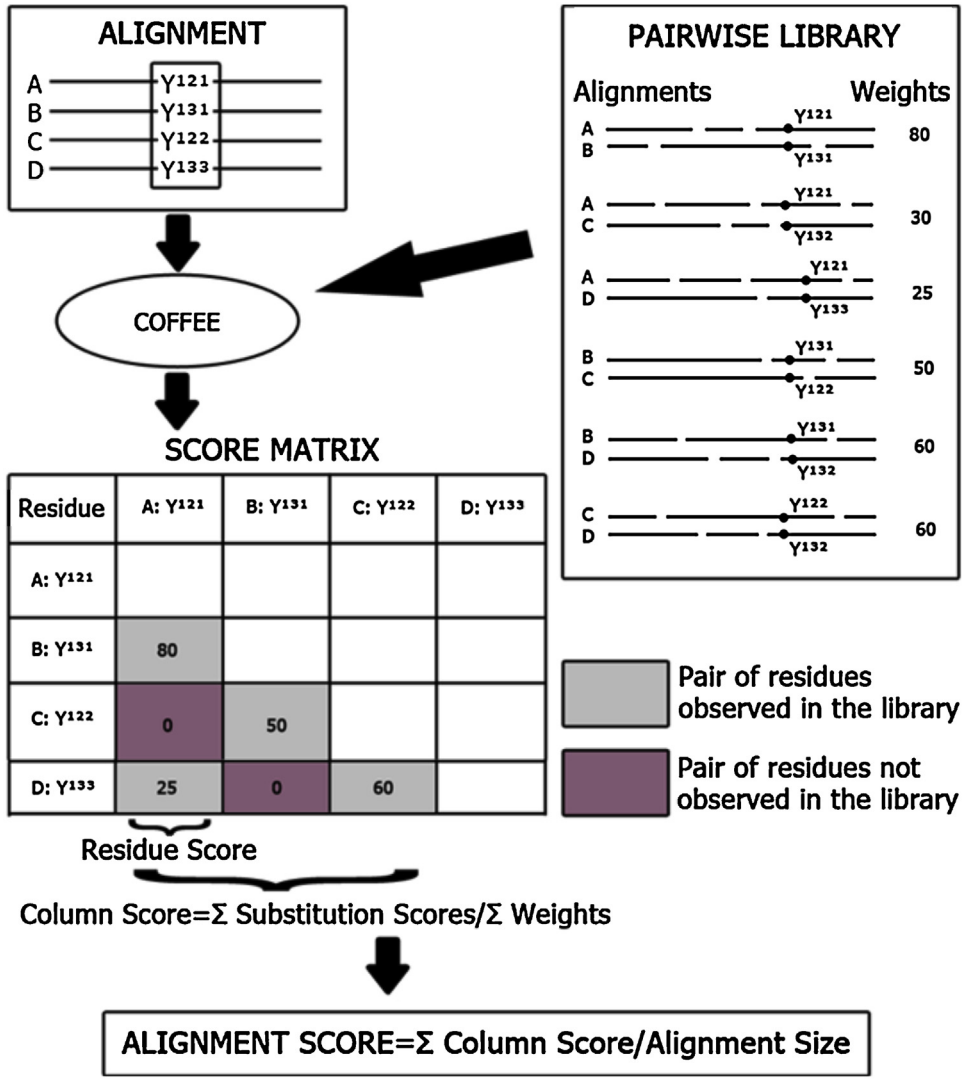

Fig. 1. COFFEE objective function scheme.