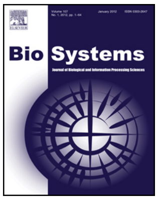
BioSystems 150 (2016) 177–189

Contents lists available at ScienceDirect

BioSystems

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



Multiple sequence alignment using multi-objective based bacterial foraging optimization algorithm

R. Ranjani Rani, Dr. D. Ramyachitra (Assistant Professor)*

Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India

ARTICLE INFO

Article history:
Received 11 April 2016
Received in revised form 18 October 2016
Accepted 18 October 2016
Available online 23 October 2016

Keywords:
Multiple sequence alignment
Multi-objective optimization
GA
ACO
ABC
PSO
GA-ABC
MO-BFO

ABSTRACT

Multiple sequence alignment (MSA) is a widespread approach in computational biology and bioinformatics. MSA deals with how the sequences of nucleotides and amino acids are sequenced with possible alignment and minimum number of gaps between them, which directs to the functional, evolutionary and structural relationships among the sequences. Still the computation of MSA is a challenging task to provide an efficient accuracy and statistically significant results of alignments. In this work, the Bacterial Foraging Optimization Algorithm was employed to align the biological sequences which resulted in a non-dominated optimal solution. It employs Multi-objective, such as: Maximization of Similarity, Non-gap percentage, Conserved blocks and Minimization of gap penalty. BAliBASE 3.0 benchmark database was utilized to examine the proposed algorithm against other methods In this paper, two algorithms have been proposed: Hybrid Genetic Algorithm with Artificial Bee Colony (GA-ABC) and Bacterial Foraging Optimization Algorithm. It was found that Hybrid Genetic Algorithm with Artificial Bee Colony performed better than the existing optimization algorithms. But still the conserved blocks were not obtained using GA-ABC. Then BFO was used for the alignment and the conserved blocks were obtained. The proposed Multi-Objective Bacterial Foraging Optimization Algorithm (MO-BFO) was compared with widely used MSA methods Clustal Omega, Kalign, MUSCLE, MAFFT, Genetic Algorithm (GA), Ant Colony Optimization (ACO), Artificial Bee Colony (ABC), Particle Swarm Optimization (PSO) and Hybrid Genetic Algorithm with Artificial Bee Colony (GA-ABC). The final results show that the proposed MO-BFO algorithm yields better alignment than most widely used methods.

1. Introduction

The protein sequence alignment problem lies at the heart of bioinformatics and computational biology that describes the way of arrangement of RNA, DNA or protein sequences and identify the regions of similarity among them. Multiple sequence alignment is one of the most prominent tasks in bioinformatics and molecular biology.

The Multiple Sequence Alignment is a method of aligning three or more sequences of RNA, DNA or proteins of similar length. In general, the input set of query sequences is expected to have an evolutionary relationship by which they share a common lineage. They are often used to evaluate the secondary and tertiary structure of protein, functional site prediction, phylogenetic analysis, Sequence Homology and conservation of Protein Domains and Motifs. Developing accurate multiple sequence alignment of different protein sequences is a difficult computational task which is an NP-Complete optimization problem (Jimin, 2008). Dynamic programming is the first approach used in bioinformatics to compare the biological sequences and find the optimal alignment. The Needleman-Wunsch algorithm is an example of Dynamic programming, which is a standard practice to align just two sequences, but it can handle only a small number of residues (Feng et al., 1984). Today the dimension of MSA problems increases significantly with their lengths and quantity of sequences. To overcome the cons of the dynamic programming method, multiple sequence alignment problems can be solved based on the concepts of Progressive method and Iterative method.

Progressive methods are efficient, but do not guarantee a global optimal solution since the sequences are added in an incorrect order in the guide tree and cannot be taken back (Phillips, 2006; Loytynoja and Nick Goldman, 2005a; Thompson et al., 1997). In order to overcome restrictions of progressive method, an iterative approach was developed by Gotoh (Gotoh, 1996). Here the iterative enhancement is done after initial alignment of progressive

* Corresponding author.
E-mail addresses: ranjaniRSR91@gmail.com (R.R. Rani), jaichitra1@yahoo.co.in (D. Ramyachitra).

http://dx.doi.org/10.1016/j.biosystems.2016.10.005
0303-2647/© 2016 Elsevier Ireland Ltd. All rights reserved.

gathering of multiple sequence alignment. This is carried out until it does not have any further improvement of alignments. This paper deals with iterative algorithms with a stochastic approach for Multiple Sequence Alignment. The rest of this paper is organized as follows: Section 2 illustrates the various related work for multiple sequence alignment methods. Section 3 describes the methods of multiple sequence alignment, Multi-objective optimization and its functions. Section 4 presents the proposed MO-BFO algorithm for multiple sequence alignment. Section 5 shows the analyses of the experiments accomplished on different datasets, comparison of results with existing methods and the implementation and discussion. Finally, Section 6 discusses the conclusion of the paper and suggests for future enhancement.

2. Related work

Among the major multiple sequence aligners, some of them employ progressive and iterative alignment approaches. They are ClustalW (Thompson et al., 1994), Clustal-X which is the GUI version of the ClustalW (Thompson et al., 1997), Clustal Omega (Sievers and Higgins, 2014), DIALIGN (Morgenstern et al., 1998), Match-Box (Depiereux et al., 1997), T-Coffee (Tree-based Consistency Objective Function for alignment Evaluation) (Notredame et al., 2000) and MUSCLE (Multiple Sequence Comparison by Log-Expectation) (Edgar, 2004). The latest generation of MSA incorporates constraint-based methods into progressive approaches like COBALT (Constraint-based alignment tool) for multiple protein sequences (Papadopoulos and Agarwala, 2007) for obtaining optimal alignment. Consistency based alignment will realign the sequences through global and local refinement methods by binding the information enclosed within the consistently aligned regions among a set of pairwise superposition. It builds a multiple alignment that is more consistent with improved pairwise alignments (Ebert and Brutlag, 2006). Few examples are ProbCons (Probabilistic Consistency-based multiple sequence alignment) (Do et al., 2005), MAFFT (Multiple Alignment using Fast Fourier Transform) (Katoh et al., 2005), Kalign (Lassmann and Sonnhammer, 2005) and Probalign (Roshan and Livesay, 2006). Align-m is a new algorithm for aligning highly divergent sequences that incorporates a non-progressive local approach which direct the results to a global alignment (Van Walle et al., 2004).

PRANK is defined as a probabilistic multiple alignment program for proteins, DNA and codon sequences and it is not intended for the alignment of much diverged sequences (Loytynoja and Goldman, 2005b). PRANKSTER is the graphical front-end of PRANK method. Consensus methods made an effort to find out the best multiple sequence alignment which results in several diverse alignments of the same set of sequences. Two commonly used methods are MergeAlign (Collingridge and Kelly, 2012) and M-Coffee (A Meta – Multiple Sequence Alignment Tool) (Wallace et al., 2006). The ClonAlign method (Layeb and Deneche., 2007) used the iterative approach, whereas PicXAA (Probabilistic Maximum Accuracy Alignment) used both the iterative and consistency based alignment approach (Sahraeian and Yoon, 2010). Sequence profiles could be redefined in probabilistic form as a profile Hidden Markov Model (HMM). Hence, they can be used instead of standard profiles in progressive and Iterative sequence alignments. Some of their examples are MUMMALS (Pei and Grishin, 2006), FSA (Fast and Statistical alignment) (Bradley et al., 2009) and MSAProbs (Liu et al., 2010).

Nowadays the problem with size of multiple sequence alignment is increased in huge volume. With the aim of handling this issue, an exceedingly promising method called stochastic optimization is used. The prominent approaches of stochastic optimization such as simulated annealing (MSASA) (Kim et al., 1994), Gibbs sampling (Lawrence et al., 1993), genetic algorithm and evolutionary algorithm are used to resolve the MSA problems (Omur Bucak and Uslanca, 2011). The genetic algorithm is used to solve the Multiple Sequence Alignment problem by optimizing simple scoring function and by using simple genetic operators (Botta and Negro, 2010). Some of the approaches based on genetic algorithm for aligning sequences are GAPM (Progressive Alignment Method using Genetic Algorithm) (Naznin et al., 2012), MSA-GA (Multiple Sequence Alignment using Genetic Algorithm) (Gupta et al., 2013), RBT-GA (Rubber Band Technique using Genetic Algorithm) (Taheri and Zomaya, 2009), VDGA (Vertical Decomposition with Genetic Algorithm) (Naznin et al., 2011) and SAGA (Sequence Alignment by using Genetic Algorithm) (Notredame and Higgins, 1996). A novel genetic algorithm for aligning multiple biological sequence was projected using the multigroup parallel and migration approach and also the novel mutation operator was designed to improve the capability to accomplish a high-quality solutions (Luo et al., 2011). Few approaches based on evolutionary algorithms are Ant Colony Optimization (ACO) (Tsvetanov et al., 2015), Particle Swarm Optimization (PSO) (Xu and Chen, 2009), M-BPSO (Long et al., 2009), Artificial Bee Colony (ABC) (Lei et al., 2010) and Genetic algorithm with Ant Colony Optimization (GA-ACO) (Lee et al., 2006). A novel multiple sequence alignment algorithm based on Ant Colony Optimization (ACO) with the divide and conquer method was developed to achieve a high quality solution. This method partition the set of sequences into numerous subsections vertically by bisecting the sequences iteratively using the ant colony optimization method (Chen et al., 2006).

Altogether the above MSA techniques are applied in single objective approach which cannot supply a set of alternative solutions that trade different solutions against each other. On the contrary, in a multi-objective optimization approach with conflicting objectives, there is no single optimal solution. The communication among different objectives gives rise to a set of compromised solutions which is known as non-dominated or trade-off. No single optimal solution is provided by multi-objective optimization (Dragan, 2002). In recent times, many works have been implemented on multi-objective based evolutionary algorithms (Abbasi et al., 2015; Soto and Becerra., 2014). Few multi-objective based alignments are MOMSA (Zhu et al., 2015), MO-SAStrE (Ortuno et al., 2013), MSAGMOGA (Kaya et al., 2014) and NSGA-II (Ortuno et al., 2012). Some efforts have been implemented for MSA to obtain accurate alignments by incorporating structural information. Two examples of structure-based alignment methods are MO-SAStrE (Ortuno et al., 2013) and 3DCoffee (O'Sullivan et al., 2004). In this paper, the combination of similarity, gap penalty, non-gap percentage and conserved blocks were used as multi-objective to obtain a non-dominated optimal alignment.

3. Methods

The multiple sequence alignment is the typical alignment between three or more biological sequences, in search of maximal similarity among them (Chellapilla and Fogel, 1999). Let $C_{1,}C_{2,}\ldots\ldots\ldots\ldots C_n$ be the input sequence strings with a minimum of three sequences. Let $\sum$ be the finite alphabet set, where the gap ('-') is not an alphabet which makes the length of the sequences to be equal to align. The multiple sequence alignment S is defined as n dimensional character array over the alphabet where $\sum' = \sum U \{-\}$. The alignment array S has n rows and each row of A_i is the alignment for string C_i when the input sequence strings are given. The $\sum$ of DNA sequences consist of 4 characters {A, T, C, G} of nucleotide and for protein sequences 20 characters {A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V} of amino acids (Abbasi et al., 2015).