Cellular Automata-Based PSO Algorithm for Aligning Multiple Molecular Sequences

Jayapriya Jayakumar, National Institute of Technology, Tiruchirappalli, India
Michael Arock, National Institute of Technology, Tiruchirappalli, India

ABSTRACT

In Bioinformatics, sequence analysis is the basic important concept that provides information for structural and functional analysis. Multiple Sequence Alignment (MSA) is a keystone problem in the sequence analysis that is used for constructing phylogenetic tree, finding motif, gene expression, etc. Basically, all biological computation issues are NP-complete problems. In this paper, a novel approach using Cellular Automata (CA) and Particle Swarm Optimization (PSO) techniques are proposed for MSA problem. Both of these techniques handle NP-complete problems very skillfully. For experimental analysis, the BaliBASE benchmark dataset and two scoring functions Sum of Pairs (SP) and Total Column (TC) are considered in this paper for calculating the similarity among the sequences. Using the Wilcoxon matched pair signed rank test the significance of the proposed algorithm (PSOCA) is explained. This algorithm is compared with PSO, genetic algorithm and the state-of-the-art techniques. The results show that the PSOCA approach yields better performance than other state-of-the-art algorithms.

KEYWORDS

Cellular Automata, Multiple Sequence Alignment, Pair-Wise Alignment, Particle Swarm Optimization, Sum of Pairs, Total Column

1. INTRODUCTION

The main goal of Bioinformatics is to understand a living cell better and its functionality at the molecular level. This can be fulfilled by the analysis of the raw molecular sequences and structural data. In sequence analysis, different concepts such as sequence alignment, motif and pattern discovery are covered. In core, sequence alignment is a fundamental problem in computational genomics. The sequence represents one or more genomes where all its length should not be compulsorily equal. Here, the sequence is nothing but the combinations of letters of the alphabet. Different types of sequences like DNA, RNA and proteins are formed by the “central dogma” of biology in which DNA is transcribed to RNA and that is translated into proteins (Xiong, J. 2006).

Sequence alignment is a concept, in which similarity or difference between the sequences is found. When two sequences are compared to align according to their similarity, it is known as pair-wise alignment. As the extension of pair-wise alignment, MSA was introduced. In this alignment, three or more sequences are aligned. Figure 1 represents the two different forms of sequence alignment. The MSA can be defined as the process that arrange n sequences where n>2 in such a way that a maximum number of residues from each sequence is matched.

The MSA reveals more biological information than any pair-wise alignments can do. Multiple Protein sequence alignment identifies many conserved and functional critical amino acid residues. The three major tasks in the MSA are scoring, creating an alignment and assessing its significance.
The traditional alignment methods can be characterized into two forms, namely exhaustive and heuristic algorithm. Exhaustive methods need extract dimensions to take all possible ways of sequence matching like dynamic programming sequence matching in pair-wise alignment. The drawback of this method is computationally intensive and handles only dataset with the small number of sequences. To overcome this problem, many heuristic algorithms have been developed. This can be classified into Progressive Alignment (PA), Iterative Alignment (IA) and Block-based Alignment (BA) types. BA method (Mohsen et al., 2011) is used to find the conserved domains and motifs. It is not done through PA and IA methods.

The PA algorithms initially start aligning most similar pairs of sequence and then continue with less similar ones. Even though PA is fast, it is not suitable for aligning sequences of different lengths. Some of the progressive type algorithms are T-Coffee (Notre dame C et al., 2000), DBClustal (Thompson J D et al., 2000), PRALINE (Simossis V A et al., 2005), where PRALINE is more sophisticated and accurate alignment program, but extremely slow in terms of time. This problem is overcome by the iterative alignment. The main idea of this is to repeatedly modify suboptimal solutions to find the optimal solution. In general, more information is gathered from global alignment than the particular divergent residues. An important advantage of Evolutionary Algorithm (EA) based approach over progressive methods is that, different fitness function can be tested without modifying the alignment procedure. There are many works proposed using EA like (Notre dame C et al., 2000; Rasmussen TK et al., 2003; Lee Z J et al., 2008; Kaya M et al., 2014). In essence, EA based methods reduced the computational time for NP-complete problems. In this proposed work, one of the evolutionary algorithms, PSO along with CA rules is used. All the algorithms for MSA are evaluated according to a particular scoring function which is the objective function for finding one of the best solutions. The scoring function, Sum of Pairs (SP) and Total Column Score (TC) are considered in most of the algorithms.

Notre dame C et al., (1996) has proposed a genetic algorithm, namely SAGA with sum of pairs as objective function using tertiary structure datasets. Notredame C et al., (2000) has proposed a progressive alignment based method using position-specific scoring scheme. Yin et al in (2002) has proposed a new approach for multiple sequence alignment using genetic algorithms (MSAGA). This algorithm uses the fitter spanning tree for sequence alignment. Rasmussen T K et al., (2003) has proposed improved Hidden Markov Model with PSO using SP as objective function. Edgar (2004) has proposed a new computer program called MUSCLE for multiple sequence alignment problem using two distance measures. For unaligned pair k-mer distance and for aligned pair the Kimura distance are employed. Lee et al., (2008) has developed a hybrid algorithm with GA (Genetic Algorithm) & ACO (Ant Colony Optimization) using SP score. Nazin et al., (2012) has developed a progressive...
Support for Dynamic Trading and Runtime Adaptability in Mobile Environments
www.igi-global.com/chapter/support-dynamic-trading-runtime-adaptability/4222?camid=4v1a

A Novel Time Series Forecasting Approach Considering Data Characteristics
www.igi-global.com/article/a-novel-time-series-forecasting-approach-considering-data-characteristics/120580?camid=4v1a