Abstract

The problem of multiple sequence alignment has been widely used in the field of bioinformatics, of which predictive structure, biological analysis, and phylogenetic simulation can be mentioned as some of its applications. A wide range of methods have been proposed to solve the problem of multiple alignment of sequences. Dynamic programming methods, heuristic methods, meta-heuristic or AI-based methods are part of the techniques used to solve the problem of multiple alignment of sequences. The most important disadvantage of dynamic programming algorithms is their very high time order, and also with these methods a limited number of sequences can be aligned. Accordingly, heuristic algorithms, such as iterative methods and progressive methods, were proposed. One of the most important limitations of heuristic methods is getting stuck in the optimal local trap. Based on this, methods based on artificial intelligence (AI) were proposed. Artificial intelligence is a branch of computer science that consists of various fields such as evolutionary algorithms, genetic algorithms, Swarm intelligence, simulated annealing, machine learning, natural language processing, and more. Each of these areas is applicable to bioinformatics, and in particular to solving the problem of multiple sequence alignments.

In this dissertation, the problem of multiple alignment of protein sequences with the help of genetic algorithm and word-to-vector algorithm (Word2Vec) has been solved. The process of implementing the algorithm as a whole is divided into two parts. In the first part, the Word2Vec algorithm is executed first. To do this, first the k-mers are extracted by the overlapping processing method, and then the Word2Vec algorithm uses the Skip-gram model to generate word embedding and forms the vector space. In the vector matrix formed, each sequence represents one row of the matrix. In the second part, with the help of a vector matrix, the distance between the sequences is measured by the distance metric, and the closest sequences to the longer sequence are identified. Then, the genetic algorithm is performed and the alignment process begins with a pattern similar to the progressive method. In other words, alignment begins with close sequences, and other sequences are added to the alignment in order of distance. The IN/DEL weighting mechanism has also been applied. It considers different weights depending on the location of IN/DEL in different areas. This system was created by performing some modifications to the Sum of Pairs scoring function. The performance of the proposed method has been investigated using BAliBASE 2.0 database. The proposed algorithm has been compared with several other algorithms, and the results show that although due to the use of the genetic algorithm, the execution time of the algorithm has increased, but the alignments created are competitive compared to the alignments obtained from other methods.

Keywords: Insertion and Deletion Mutations, IN/DEL, Sequence Alignment, Artificial Intelligence, AI, Word2Vec, Genetic Algorithm, GA

University of Zabol

Faculty of Basic Sciences

Investigation of IN/DEL Weighting in Aligning of Protein Sequences Using Artificial Intelligence

Supervisors

Dr. Ali Maghsoudi

Dr. Mohammad Allahbakhsh

Advisor

Mohammadreza Pourmir

By

Hamidreza Hosseini

Date

June, 2020