Manish Kumar /J. Pharm. Sci. & Res. Vol. 7(6), 2015, 314-318

ISSN:0975-1459
Journal of Pharmaceutical Sciences and Research
www.jpsr.pharmainfo.in

A Comparative Study on the Alignment Quality of Multiple Protein Sequences

Manish Kumar

CSE Department, Indian School of Mines, Dhanbad-826004, India.

Abstract

Aim: The objective of this study is to infer into the protein sequences in order to solve the multiple sequence alignment problem of protein sequences using genetic algorithm optimizing technique.

Methods: In this work, a genetic algorithm based approach has been presented and used accordingly. Different genetic operators along with a fitness function is proposed so as to obtain a optimal quality solutions for protein sequences.

Results & Discussions: Results based on various parameters have been recorded and analyzed over a set of clusters orthologous groups of proteins and were compared with the results obtained with other alignment algorithms, e.g. Clustal W and Central Star algorithms. The obtained results show the superiority of the proposed technique as it achieves betters solution in terms of scores, when compared with the methods mention above.

Keywords: Bioinformatics; Crossover Operator; Genetic Algorithm; Multiple Sequence Alignment; Mutation Operator.

INTRODUCTION

Bioinformatics is an inter-disciplinary subject from biology, mathematics and computer science, and also an important frontier of today's life science and natural sciences. With the advent of molecular biology and of genome research, biological data has grown rapidly and reached huge volumes. Since there are close relationship and association among these datasets, making full use of these data and getting the useful information by the analysis and processing, revealing the connotation of these data are the primary challenges of bioinformatics today [1]. One of the primary concerns in this field is analysis and interpretation as well as prediction of DNA molecular structures and protein structures [2]. In addition, other related open research areas are alignment problem (both single and multiple), drug discovery, phylogenetic tree generation, gene regulatory network etc.

The multiple sequence alignment [3] of protein sequences or DNA sequences has become one of the most important tools in the modern molecular biology, especially with the implementation of the "Human Genome Project", more and more sequences have been obtained and need to do the insightful analysis. It is a method of arranging the primary sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences are generally represented in terms of rows and columns of a matrix. Gaps are inserted between the amino acid residues so that residues with identical or similar characters are aligned in successive columns [4]. If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations and gaps as indels (that is, insertion or deletion mutations) introduced in one or both species in the time since they diverged from one another.

To determine the level or score of similarity, many methods are currently available. Two most popular optimal alignment algorithms are the Needleman- Wunsch algorithm [5] and the Smith-Waterman algorithm [6]. The Needleman-Wunsch algorithm is one of the classic methods in finding an optimal global alignment of two sequences by maximizing the number of matching DNAs and minimizing the number of gaps. The Smith-Waterman algorithm is similar to the Needleman-Wunsch algorithm except that it enables local sequence alignment. Instead of aligning the entire length of two protein sequences, this algorithm finds the region of highest similarity between two sequences. This is potentially more biologically relevant due to the fact that the ends of proteins tend to be less conserved than the middle portions, leading to higher mutation, deletion, and insertion rates at the ends of the protein.

BLAST (Basic Local Alignment Search Tool) is an approximate algorithm that is very popular in biological research, especially genetic research and bioinformatics. The algorithm finds the highest scoring locally optimal alignments between a query sequence and a database of sequences [7]. The tool is designed to discover all of the similar sequences in the database and create a statistical interpretation to enable the user distinguish a particular DNA or protein from random background hits. There is a modification to BLAST, called Gapped BLAST. It was developed to use "twohit" approach in which a word can be followed by a second word which is within a certain gap threshold [8]. These matches are then extended using a matrix in all directions until the score drops down to a certain percentage threshold of the highest score computed. However, Gapped BLAST may also yield suboptimal alignment results since when it performs dynamic programming at the end, the best alignment may lie outside of the range that it has defined.

CLUSTAL is a program for multiple sequences alignment which uses the "progressive"approach by Feng-Doolittle. ClUSTALW is the latest version of this series except the X version which provide a graphic interface. The W means "weighting", it can provide the "weights" to the sequences and the program parameters. ClUSTAL W [9] improves the sensitivity of the progressive multiple sequences alignment

through three additional heuristics including sequence weighting, position specific gap penalties and weight matrix choice. However, most of the existing algorithms do not allow the users to easily access and modify scoring values. These algorithms are not focused on information specific purposes.

All the MSA construction methods studied in the literature review are generally evaluated using one or more alignment benchmarks, for example, BAliBASE [10], OxBench [11] or PREFAB [12], and it is clear that this benchmarking has had a positive effect on their development [13]. Most of the widely used MSA benchmarks were compared in [14] and are also discussed in [15]. The use of objective benchmarks leads to a better understanding of the problems underlying poor performance, by highlighting specific weak points or bottlenecks.

In this study, a set of clusters orthologous groups of proteins has been used as a benchmark dataset to evaluate and compare the results of the proposed scheme.

In this work, a novel approach has been designed to solve the MSA problem of protein sequences though genetic algorithm by defining a fitness score to calculate the quality of the aligned sequences. Protein sequence alignment is the task of identifying evolutionarily or structurally related positions in a collection of amino acid sequences. Although the protein alignment problem has been studied for several decades, many recent studies have demonstrated considerable progress in improving the accuracy or scalability of multiple and pairwise alignment tools, or in expanding the scope of tasks handled by an alignment program. Genetic algorithms are powerful methods of optimization and used for successfully indifferent problems. Their performance is depends on the encoding scheme and the choice of genetic operators especially, the selection, crossover and mutation operators. In this paper, an enhanced genetic algorithm has been developed and adapted to the permutation presentations that can be used in a large variety of combinatorial optimization problems.

Here, the proposed method has been compared by Clustal W and Central Star algorithms [16, 17] by creating a unified experiment environment in which every program will use the same input and output and then compare the results on both the accuracy and quality. In order to insure the quality of the comparison, this study chooses to run a full experiment on a set of clusters orthologous groups of proteins so that assessment based on quality alignment of each test case can be made. All the programs are processed with default parameters which is most commonly used by the normal users.

MATERIAL AND METHOD

Population initialization

Each organism in the GA consists of a candidate alignment. For creating the initial alignment, the organisms of the initial population are generated from pairwise alignments of all the sequences. Initially, all global pairwise alignments between the sequences are computed with dynamic programming using the Needleman-Wunsch algorithm [18]. For each sequence one of the pairwise alignments corresponding to that sequence is randomly selected to form the organism. At the beginning of the sequence, a randomly defined number of gaps is placed (offset).The number of gaps is an integer that varies between 0 to 20% of the total size of the sequence. Once an organism is constructed, the objective function is defined and an initial fitness value is assigned to the organism. With this approach, the initial population starts with a high mean of fitness.

Fitness evaluation

In this sub section, a formal definition of the sum-of-pairs of multiple sequence alignment is introduced which is used as a tool to calculate fitness.

Commonly used measure for evaluating the accuracy of MSA programs is to compute *SPS* and *CS* . By counting aligned residue pairs *SPS* can be calculated. It determines MSA tools ability to align some, if not all, of the sequences in an alignment. Let us consider an alignment of N sequences comprise M columns. The c^{th} column can be assigned as A_{c1}, A_{c2},…., A_{cN}. For each pair of residues A_{cj} and A_{ck}, it is defined S_{cjk} such that $S_{cjk} = 1$, if A_{cj} and A_{ck} are in the same column of reference alignment. The score for cth column (S_c) can be defined as follows.

$S_c = \sum_{j=1}^{N} \sum_{J \neq K}^{N} {}_{K=1} S_{cjk}$

For full alignment the sum of pair score can be computed as:

$SPS = \sum_{c=1}^{M} S_C \quad / \sum_{i=1}^{C_r} S_{rc}$

C_r denotes number of columns and S_{rc} represents the score of the c^{th} column in reference alignment.

The ability to align all the columns of a given sequences by a MSA tool is determined by column score or match column score. It is calculated by dividing the total number of matched columns between test and reference alignments with the total number of “considered” columns in the test alignment. Here, for the experimental analysis it is considered that, C_c = 1 if a column of a (test) alignment matches with the column of reference alignment otherwise it is zero.

$CS = \sum_{C=1}^{M} C_c \, / M$

Crossover operation

In genetic algorithms, crossover [19] is a genetic operator used to vary the programming of a chromosome or chromosomes from one generation to the next. It is analogous to reproduction and biological crossover, upon which genetic algorithms are based. Cross over is a process of taking more than one parent solutions and producing a child solution from them.

This operator defines a cut point at a random chosen point in the alignment. After that, swapping of parents has been done in order to produce offspring. As the parents involved in crossover operation are of different lengths, so the resulting alignments are filled with gaps as Fig. 1 demonstrates.

Mutation Operation

A mutation operator [20] is defined and applied for the proposed approach with GA. Mutation operator randomly flips some of the bits in a chromosome. For example, the string 00000100 might be mutated in its second position to yield 01000100. Mutation can occur at each bit position in

a string with some probability, usually very small (e.g., 0.001).
Example of mutation operator used in this study.
Here, A is replaced with G and C with T and vice versa. For example,
the parent:
G | ***C T A A T*** | *A*
produces an offspring
G |***T A G G A*** | *A*
Here ' |' indicates the mutating point.

New generation
In this paper, tournament selection [21] is implemented for selection operator. This selection scheme is to determine which alignments in the selection pool are to become parents for the next generation in the algorithm. In the selection process, each alignment is compared with 50 opponents that are randomly selected from the selection pool. For each comparison in which the fitness of the alignment is equal to or higher than that of the opponent, the alignment receive a win. The alignments with the highest number of wins are selected to be the parent alignments for the next generation.

Termination condition
The algorithm is made to be terminated after fixed number of generation i.e. 50 or after reaching desire fitness value.

RESULTS AND DISCUSSION

The proposed approach is implemented in C. All tests have been fulfilled on a PC with an Intel i7 core 2.53 GHz processor and 4GB RAM. The experiments for each datasets are processed with default parameters which is most commonly used by the normal users. In order to evaluate the proposed approach, the experiment is carried out with a set of clusters orthologous groups of proteins. For each of the experiment, alignments were performed both with the proposed method as well as with the other methods described in the literature stated earlier. Performance, in terms of both fitness score and match column, are summarized for several of the experimental runs. The accuracy of an alignment resulting from an MSA tool is usually measured by two measurements: Fitness score and Match Column (MC). MC is the number of correctly aligned columns to the number of columns in the reference alignment and Fitness score is the number of correctly aligned residue pairs to the number of residue pairs in the reference alignment. Table 2 indicates that for the dataset COG1510 the proposed method didn't give optimal result for fitness score as the outcome of the fitness score is less than Clustal W method.

Table 1: Data sets

ID	Number of sequences	Mean Sequence length(min, max)	Similarity
COG2178	3	211(196,222)	0.1760
COG2157	4	72(57,78)	0.1500
COG1476	5	71(66,79)	0.2588
COG2097	6	96(81,113)	0.1043
COG1510	6	170(152,158)	0.0213
COG0219	9	158(151,166)	0.1404

Table 2: Comparative result between different methods

ID	Clustal Walgorithm		Central – Star algorithm		The proposed approach	
	Fitness score	Match column	Fitness score	Match column	Fitness score	Match column
COG 2178	384	41	362	50	**402**	**56**
COG2157	499	12	438	13	**545**	**18**
COG1476	1657	22	1620	22	**1874**	**26**
COG2097	1781	12	1421	13	**2154**	**15**
COG1510	**1650**	4	395	4	1457	**9**
COG0219	9734	24	7445	27	**11235**	**32**

Above table shows the comparison results, by which one can conclude that the proposed algorithm has better results among these approaches. The bold faced data represents the best scores among the methods.

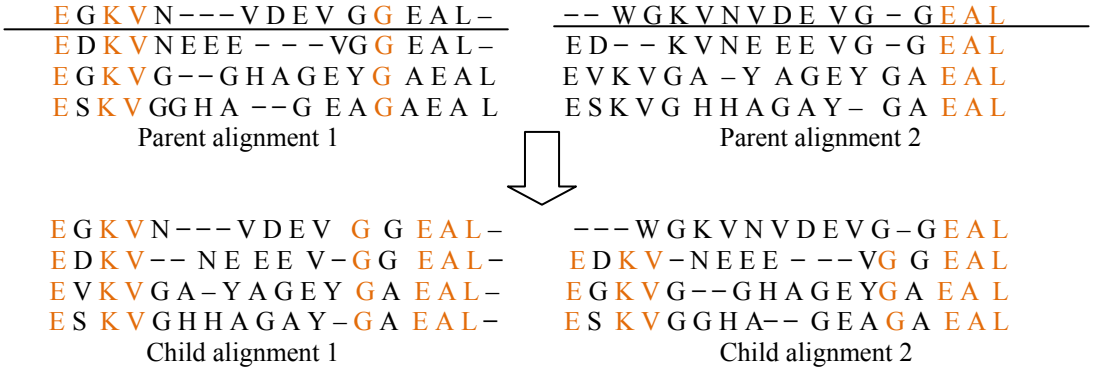

Fig. 1: Proposed crossover operator

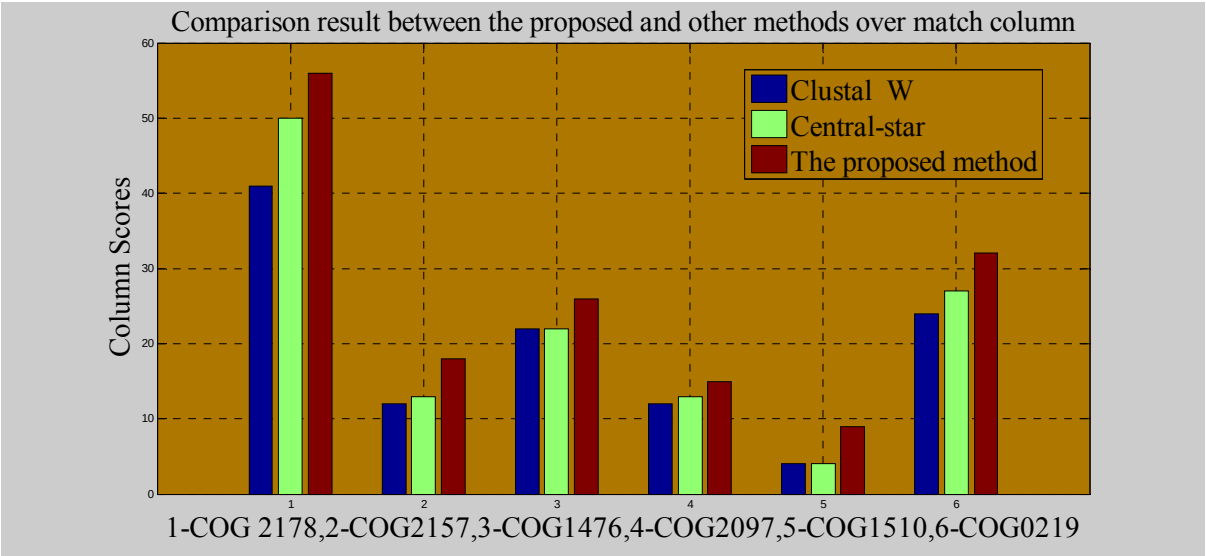

Fig.2: Bar graph comparison result of match column scores between proposed and other methods.

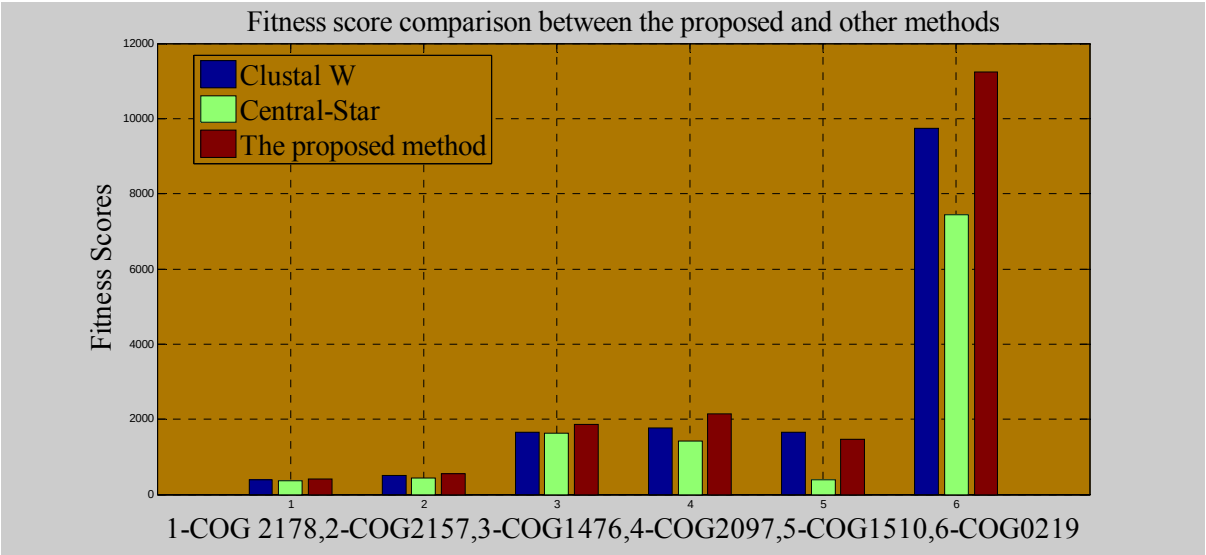

Fig.3: Bar graph comparison result of fitness scores between proposed and other methods.

CONCLUSION

The currently used techniques for multiple sequence alignment are characterized by great computational complexity, which prevents the techniques from wider use. The research reported in this paper is aimed to develop a new technique for efficient multiple sequence alignment. Genetic algorithms are stochastic approaches for efficient and robust search. Therefore, a new technique based on genetic algorithm has been proposed, where the crossover and mutation operators are efficiently used in order to get reliable result. It has been show that, how these operators can affect the optimal alignment quality of the sequences. To test the feasibility of the proposed approach, it has been compared with Clustal W and Central Star algorithms over a set of clusters of orthologous groups of proteins sequences. Compared to these algorithms, the proposed approach improves the mathematical and biological quality for many sequences with different characteristics. The results provide clear empirical evidence that the proposed method outperformed almost all the test cases considered in the experimental study. It can also be concluded that the results are promising and articulate the performance of the presented approach.

REFERENCES

[1] T sui, S.K.-W., High-throughput DNA sequencing and bioinformatics: Bottlenecks and opportunities, IEEE International Conference on Granular Computing, 2009;4 .

[2] Wang W; Liu J, Distinguishing single-stranded and double-stranded DNA binding proteins based on structural information, IEEE International Conference on Bioinformatics and Biomedicine 2013;612.

[3] Hamidi, S.; Naghibzadeh, M.; Sadri, J., "Protein multiple sequence alignment based on secondary structure similarity," International Conference on Advances in Computing, Communications and informatics, 2013,1224-29

[4] Lo¨ytynoja A, Goldman N "Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis."*Science*, 2008, Vol:320: 1632–635.

[5] S. B. Needleman and C. D. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins", *Journal of Molecular Biology*, 1970, 48, 443-453.
[6] R. Smith and M. S. Waterman, "Comparison of Biosequences", *Advanced Application Mathematics*, 1981, 2,482-489.
[7] S. Altschul, A. Madden, and W. Zhang, "Gapped BLAST and PSI-BLAST – A New Generation of Protein DB Search Programs", *Nucleic Acids Research*, 1997, 25(17).
[8] http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html.
[9]. Mahram, A.; Herbordt, M.C., FMSA: FPGA-Accelerated ClustalW-Based Multiple Sequence Alignment through Pipelined Prefiltering, IEEE 20th Annual International Symposium on Field-Programmable Custom Computing Machines, 2012,177-183.
[10] Thompson JD, Plewniak F, Poch O (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics 15:87–88.
[11] Raghava GP, Searle SM, Audley PC, Barber JD, Barton GJ (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy.BMC Bioinformatics 4: 47.
[12] Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
[13] Dessimoz C, Gil M (2010) Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol 11: R37.
[14] Blackshields G, Wallace IM, Larkin M, Higgins DG (2006) Analysis and comparison of benchmarks for multiple sequence alignment. In Silico Biol 6: 321–339.
[15] Aniba MR, Poch O, Thompson JD (2010) Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. Nucleic Acids Res 38:7353–7363.
[16] Thompson J. D., Higgins D. G., and Gibson T. J., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Research, 1994 :22. 4673-4680.
[17] Feng D. F. and Doolittle R. F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees, Journal of Molecular Evolution, 1987: 25, 351- 360.
[18] Needleman S. B. and Wunsch C. D., A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol., 1970, 48: 3, 443–53.
[19]. Abdoun O. and Abouchabaka J, A Comparative Study of Adaptive Crossover Operators for Genetic Algorithms to Resolve the Traveling Salesman Problem. IJCA, 2011: 31,11.
[20] Abdoun O, Jabouchabaka J, Tajani C .Analyzing the Performance of Mutation Operators to Solve the Travelling Salesman Problem International Journal of Emerging Sciences , 2012,2(1), 61-77
[21] Hong Y; Kwong S; Ren Q; Wang X, A comprehensive comparison between real population based tournament selection and virtual population based tournament selection, IEEE Congress on Evolutionary Computation, 2007, 445-452.