Research Article

A Heuristic Technique for Alignment of Multiple Biological Sequences Using Enhanced Evolutionary Algorithm

Manish Kumar*
Department of Computer Science and Engineering, Indian School of Mines, Dhanbad, Jharkhand, India.
***Corresponding author's E-mail:** manishkumar@cse.ism.ac.in

Accepted on: 18-10-2015; **Finalized on:** 30-11-2015.

ABSTRACT

An efficient recursive approach is proposed in this paper that would not only find the multiple sequences alignment for protein sequence but also provides means for consideration of gaps between them. MSAs are usually scored with the Sum-of-Pairs (SP) function and the Match Column (MC) function, but exact SP and MC for MSA is known to be NP-Hard. Therefore in this paper, a heuristic method is used to solve MSA problem using genetic algorithm. Three different operators were proposed, one type of selection operator, one types of crossover operators and one type of mutation operator for feasible alignment of protein sequences. The input variables (e.g. Strands) of the program are user dependant and internal calculations are performed in recursive fashion to add intelligence to the input Strands. Experimental results of benchmarks from the BAliBASE 3.0 shows that the strategy adopted in the study is able to obtain better results, when compared to the traditional MSA tools.

Keywords: Bioinformatics; Multiple Sequence Alignment; Genetic Algorithm; Genetic Operators.

INTRODUCTION

Multiple sequence alignment[1] is an essential pre-requisite in molecular sequences analysis. This has lead to the development of different software tools. MSA can be used in Phylogenetic analysis to trace the path of evolution. The most general purpose of multiple sequence alignment is to find highly conserved region or embedded patterns. Patterns / Motifs are well conserved regions of sequence generally organized around one or two very highly conserved residues.

Multiple Sequence Alignment (MSA) is identified as one of the challenging tasks in bioinformatics which belongs to a class of hard optimization problems called combinatorial problems. Multiple sequence alignment allows comparison of sequences by simultaneously aligning set of sequences. The main problem in MSA is its exponential complexity with the considered input data set. These alignments may be used to identify profiles or hidden models that may be used to acquire knowledge for distantly related members of the family sequences, newly discovered sequences, and existing sequence databases.

MSA is an optimization problem which exhibits a great temporal and space complexity. Therefore, several methods were proposed which can be grouped in three great classes[2]. Initially, solution was presented by Needleman algorithm[3]. The second class contains methods based on a progressive approach[4]. The progressive methods are simple, fast and generally give alignments of good qualities. However, their major disadvantage is the problem of the local minima and consequently they can lead to poor quality solutions. The first stochastic iterative algorithm proposed in the literatures uses an algorithm of simulated annealing[5]. However, this algorithm is very slow and it is inappropriate to be used as improver[2]. Later, several other iterative algorithms which use various strategies like Genetic Algorithms (GAs) were proposed. Concerning the deterministic iterative methods, they involve extracting the sequence one by one from multiple alignments and realigning them to the remaining sequences. The major disadvantage of iterative method is their high execution time.

Thompson[6] described that traditional algorithms such as Clustal W are known to be very successful when the number of average length is low and similarity is high. They also said that Zhang and Jorong Tzong horng developed a genetic algorithm for MSA. According to Thompson[6] no single alignment procedure can be expected to construct biologically reasonable alignment in all possible situations. Some authors were also presented a novel algorithm with self-organizing neural network for MSA. Self-organizing NN as local optimization like classification is embedded into genetic algorithm to keep away from local optima.

Chen and others[7] presented a new method for multiple DNA sequence alignment using genetic algorithms and divide-and-conquer techniques to choose optimal cut points of multiple DNA sequences. Their experimental results show that their method is better than the previous methods presented in various other literatures[8] for dealing with multiple DNA sequence alignment from the viewpoints of the scores and the match column. But as per Chen[7], unfortunately degree of similarity between two sequences in the fitness functions not to be calculated by aligning two sequences to the "left" by inserting the symbols "-" to the right-hand side of the shorter sequence.

Referring to other literature studies and according to Omar[9] multiple sequence alignment relies very much on optimization algorithms. The combination of genetic algorithm and simulated annealing was a way that can be used to solve MSA assignment. As per Peng[10] genetic algorithm will try to find new region of feasible solution while simulated annealing will act as aligning improver. Simulated annealing also helps to prevent local minima problem compared to the dynamic programming. But literature studies states that, further tests need to be carried out to prove that the use of SA can produce better results. Pengfei[11] in their experiments showed that GA itself is sufficient to solve the problem. However, this is not true since SA can avoid local minima. There are other aspects of the system that need to be improved. The coding schedule needs to be tested thoroughly to get better results. Another factor is due to the gap insertion process, where a new operator can increase the performance of the system.

Otman[12] in their research explained that the mutation operators have an important role in introducing new patterns in the population. Also, the number of generations and other scoring matrixes can have influence in the results for different datasets, but in this study as per them they kept these parameters unchanged in order to establish a similar environment for all test configurations. A straightforward development of their investigation is to determine a set of rules that can evaluate the evolution of the population and that choose the correct type of operator at a given time.

With consideration to all the above facts about MSA problem, an approach has been made in this research to increase the alignment quality of multi biological sequences by enhancing the genetic parameters of genetic algorithm. Here, an appropriate development is made right from the initial generation of population to the selection of individuals for crossover and mutation. The traditional genetic operators such as the crossover and mutation were also modified in order to increase the quality of solution for alignment problem. Later on, the alignment so obtained were compared over the sum of pair score and match column with tools like Mafft[13], Probcons[14], Muscle[15], Clustal W[16], t-Cofee[17] and MSACompro[18].

MSA tools using genetic algorithms and simulated annealing technique have also been developed in an attempt to seek a more optimized MSA. In the genetic algorithm based method, a set of already generated multiple sequence alignments is taken and divided into many fragments. These fragments are then rearranged to obtain a more optimal solution. The process is repeated until it converges where the objective function is maximized.

Though many MSA tools have been developed so far, Clustal-W continues to be the most widely used MSA tool. T-COFFEE and PROBCONS are best tools as far as accuracy is concerned but are not scalable beyond 100 sequences. Therefore while aligning more than 100 sequences MAFFT and MUSCLE give the best performance in terms speed and accuracy. Most of these MSA tools have been designed considering protein sequence alignment but they can be fairly used for DNA and RNA sequences also.

The rest of the paper is organised as follow. The next section introduces the concepts underlying the research work with detailed discussion on the proposed approach. Followed by section which explains about the detailed results over standard datasets, along with the experiments setup required in order to validate and observe the results. Finally, the concluding section presents the final consideration.

MATERIALS AND MEHTODS

Initial Population Generation and Selection

The selection operator determines which chromosomes will survive in each generation. Therefore, in this process the combination of Roulette wheel selection and Elitism is implemented. The selection of fittest chromosomes within each generation is guaranteed by Elitism.

The initial generation of number of chromosome is made to be generated by the choice given by the user. For performing the crossover and mutation operation, the chromosomes are selected using the Roulette Wheel selection scheme. After the selection process is over, the chromosomes are subjected for crossover and mutation operation and the best chromosome from the current population is selected and saved as elite. Based on the best fitness score the elites are replaced.

Fitness

The sum-of-pairs score (SPS): is calculated so that the score increases with the number of sequences correctly aligned which is used to determine the extent to which the programs succeed in aligning. The SPS score is defined as below:

Considering a test alignment of size *NxM,* and a reference alignment of size NxM_r, where *N* is the number of sequences, and *M,*M_r are the number of columns in the test and reference alignment accordingly and A_{i1} A_{i1}, …, A_{iN}, is the i^{th} column in the alignment, for each pair of residues A_{ij} and A_{ik} we define p_{ijk} = 1 if residues A_{ij} and A_{ik} are aligned with each other in the reference alignment, otherwise p_{ijk} = 0. The score S_i for the i^{th} column will be the sum of p_{ijk} for all pairs of symbols in this column:

$S= \sum_{j=1,j\neq 1}^{N} \sum_{k=1}^{N} P_{ijk}$

Similarly S_{ri} is the score S_i for the i^{th} column in the reference alignment.

The SPS score for the test alignment is:

$SPS = \sum_{l=1}^{M} S_i / \sum_{i=1}^{Mr} S_{ri}$

The Column score (CS)

The Column score (CS): Considering a test alignment of size *NxM,* and a reference alignment of size NxM_r, where

N is the number of sequences, and M, M_r are the number of columns in the test and reference alignment accordingly: the score C_i = 1 if all the residues in the column are aligned in the reference alignment, otherwise C_i = 0

The CS score for the test alignment is then:

$CS=\sum_{i=1}^{M} C_i$ / M

Since, the two scoring systems have been implemented successfully in the program BAlibase called BAliscore which takes as input a test alignment and a reference alignment in MSF format, in this paper BAliscore has been used to estimate the quality of the test alignment.

Crossover Operator

The crossover operator combines the genes of two or more parents to generate better offspring. It is based on the idea that the exchange of information between good chromosomes will generate even better offspring. The effect of the crossover operator can be studied from two different points of view: at chromosome level and at gene level.

The proposed crossover operators described in fig. 1 is used for successful alignment of biological sequences. It can be seen that, two parents a and b were divided from a random selected point into a1, a2 and b1, b2. Then, a2 and b2 were further divided to form a3, a4 and b3, b4. At the last, (a1+b3+a4) and (b1+a3+b4) were combined which can be seen in the figure to obtain the optimal alignment of sequences.

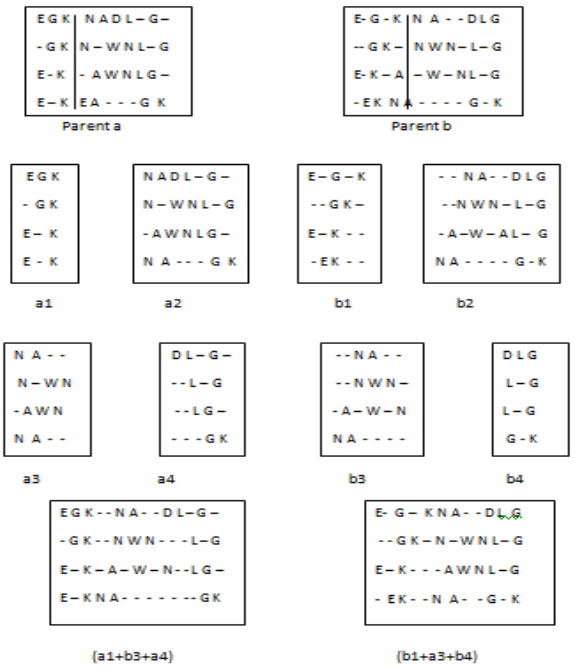

Figure 1: The proposed crossover operator

Mutation

The main aim of mutation operator is to slightly alter the parent to introduce new genetic information. The proposed scheme of mutation operator works as follow. First, a shorter segment from the parent is chosen at random, which is limited to *8≤ l≤ 90*. Then the chosen segment is divided into two groups from a random chosen position. In each group, the column consisting of only gap character were removed and the Myers-Miller algorithm is used to re-align these two groups to a segment of alignment. Finally, the new segment is connected to two terminal segments of the parent to complete the offspring. Now, if the newborn child is different compared to previously generated children then, it will be put into the new generation, otherwise, it will simply be discarded. As, the length of the short segment is limited to *8≤ l≤ 90*, the computational time for the mutation is bound by a constant, not dependent on the length of sequences of the problem.

Termination Condition

The termination conditions used for the experiment are as follows:

In the experimental study, the results were tasted on maximum 70 iterations (generations), and hence made the experiment to be terminated after reaching 70 iterations, as there is negligible amount of improvement in the alignment quality.

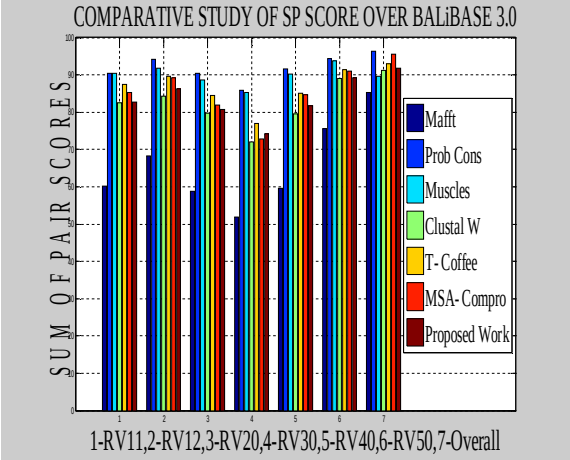

Figure 2: Bar graph comparison result of SP scores between the proposed method and MSA tools.

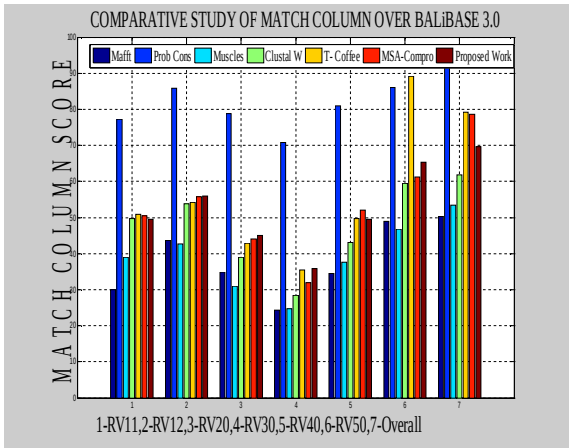

Figure 3: Bar graph comparison result of MC scores between the proposed method and MSA tools.

RESULTS AND DISCUSSION

To evaluate and validate an MSA tool, a standard benchmark database of reference alignment is used. For comparing the presented work to some other MSA approaches, the BAliBASE 3.0 is used. BAliBASE 3.0 is the most commonly employed protein alignment benchmark. Reference 3 consists of up to 4 sub-groups, with less than 25% residue identity between sub-groups. It contains five main categories. All categories, except for category 4 (RV40), have two different sets: one consists of full length sequences and the other has the homology regions of those sequences. Reference 1 shares sequences with a similar length. It has two subcategories: RV11 and RV12. RV11 contains 76 files which have very distant sequences sharing less than 20% identity. RV12 contains 88 files with sequences that share 20 - 40% identity. Reference 2 (RV20) has sequences with more than 40% identity and also some orphan sequences with less than 20% identity with others. RV20 contains 82 files. RV30 has families with 40% identity, but 20% identity is shared among them. It contains 60 files. RV40 has sequences with N/C terminal extensions and 49 files. RV50 has sequences with long insertions and contains 32 files.

In the proposed approach, the quality of the alignments was measured by considering Sum-of-Pair scores and Match Column (MC) scores. MC is the number of correctly aligned columns to the number of columns in the reference alignment and SP is the number of correctly aligned residue pairs to the number of residue pairs in the reference alignment. The accuracy of the presented approach is compared to the Mafft, ProbCons, Muscle, Clustal-W, T-Coffee and MSACompro MSA tools and presented in the form of graph in figure 2 & 3.

The proposed approach is implemented using C language. 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 the parameters which is most commonly used by the normal users. The population size was established to 100 individuals and the maximum number of generations was 70 with a crossover probability of 0.7%, mutation rate of 0.01% for the experiment.

Table 1 and 2 indicates that the proposed schemes is able to provide a better solutions for sum of pair score and match column score in comparison to other commonly know MSA tools. Only for dataset RV20 for sum of pair and RV40 for MC scores the proposed approach was unable to perform better.

Table 1: A comparative result of SP scores between the proposed method and MSA tools.

MSA TOOLS	RV11	RV12	RV20	RV30	RV40	RV50	Overall
Mafft	60.13	90.5	90.49	82.52	87.36	85.30	82.71
ProbCons	68.19	94.25	91.88	84.23	89.58	89.18	86.21
Muscle	58.73	90.45	88.72	79.76	84.4	81.99	80.67
Clustal-W	51.83	85.92	85.21	72.05	76.99	72.73	74.12
T-Coffee	59.53	91.54	90.18	79.53	85.13	84.62	81.75
MSACompro	75.54	94.37	**93.76**	89.09	91.34	91.06	89.19
Proposed Work	**85.34**	**96.27**	89.71	**91.26**	**92.89**	**95.47**	**91.82**

Table 2: A comparative result of MC scores between the proposed method and MSA tools.

MSA TOOLS	RV11	RV12	RV20	RV30	RV40	RV50	Overall
Mafft	30.02	77.21	38.87	49.63	50.88	50.58	49.53
ProbCons	43.58	85.82	42.64	53.73	54.18	55.83	55.96
Muscle	34.85	78.73	30.92	39.03	42.82	44.14	45.08
Clustal-W	24.36	70.75	24.62	28.35	35.41	31.97	35.91
T-Coffee	34.44	80.86	37.57	42.99	49.73	52.05	49.60
MSACompro	48.94	86.11	46.87	59.52	**89.15**	61.23	65.30
Proposed Work	**50.34**	**94.47**	**53.41**	**61.75**	79.15	**78.57**	**69.61**

CONCLUSION

MSA tools that can guide the construction of accurate phylogenetic trees are the need of the hour, as this can cater for many applications in the future. An MSA tool which does not depict the correct biological information is of no use. Genetic algorithm (GA) is one of the important and successful approaches in multiple sequences alignment (MSA) problem. In this paper, an improved GA method has been developed, which can search the solution space in a very efficient manner. With the help of different proposed genetic operators, different protein sequences were aligned successfully. The experimental results show that the improved approach presented here can obtain a better result

compared with traditional approach in aligning multiple protein sequences. The future work will focus on new representations of sequences that help in improving the accuracy and speed of the MSA tools.

REFERENCES

1. 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, 1224-1229.
2. Notredame C., Higgins D.G. SAGA: sequence alignment by genetic algorithm, Nucleic Acids Research, volume 24(8), 1996, 1515–1524.
3. NeedIman S.B., Wunsch C.D. A general method applicable to the search for similarities in the amino-acid sequence of two proteins. Journal of Molecular Biology, 48, 1970, 443-453.
4. Feng D., Doolittle R. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol., 25, 1987, 351-360.
5. Kim J., Pramanik S., Chung M.J. Multiple sequence alignment using simulated annealing. Computer applications in bioscience, 10, 1994, 419-426.
6. Thompson J. D., Higgins D. G., 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 Res., 22, 1994, 4673–4680.
7. Shyi-Ming Chen, Chung-Hui Lin, Shi-Jay Chen. Multiple DNA Sequence Alignment Based on Genetic Algorithms and Divide-and-Conquer Techniques, International Journal of Applied Science and Engineering. 3(2), 2005, 89-100
8. Hirschberg, D. S. Algorithms for the longest common subsequence problem. Journal of the ACM, 24, 1997, 664-675.
9. Omar M. F., Salam R. A., Abdullah R., Rashid N. A. Multiple Sequence Alignment Using Optimization Algorithms, International Journal of Computational Intelligence, 1, 2005.
10. Peng Y; Dong C; Zheng H. Research on Genetic Algorithm Based on Pyramid Model, 2nd International Symposium on Intelligence Information Processing and Trusted Computing, 2011, 83-86.
11. Pengfei G; Xuezhi Wa; Yingshi H. The enhanced genetic algorithms for the optimization design, 3rd International Conference on Biomedical Engineering and Informatics, 7, 2010, 2990-994.
12. Otman A, Jaafar A, Chakir T. Analyzing the Performance of Mutation Operators to Solve the Travelling Salesman Problem" International Journal of Emerging Sciences, 2(1), 2012, 61-77.
13. Katoh K; Kuma K; Toh H; Miyata T. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. Nucleic Acids Res., 33, 2005, 511-518.
14. Chuong B. D; Michael B; Serafim B. PROBCONS: Probabilistic Consistency-based Multiple Alignment of Amino Acid Sequences. American Association for Artificial Intelligence, 2004, 703-708.
15. Edgar R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res., 32(5), 2004, 1792–1797.
16. Thompson J.D; Higgins D.G; 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 Res., 22, 1994, 4673–4680.
17. Notredame C; Higgins D.G; Heringa J. T-coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol., 302, 2000, 205–17.
18. Deng X; Cheng J. MSACompro: protein multiple sequence alignment using predicted secondary structure, solvent accessibility, and residue-residue contacts. BMC Bioinformatics, 2011, 12: 472.

Source of Support: Nil, **Conflict of Interest:** None.