Research Journal of Applied Sciences, Engineering and Technology 4(21): 4404-4407, 2012
ISSN: 2040-7467

Submitted: April 18, 2012 Accepted: May 14, 2012 Published: November 01, 2012

A Refined MSAPSO Algorithm for Improving Alignment Score

[1]K. Arulmani, [2]M. Guru Prasad, [2]R. Hariharan and [2]N. Sivasankaran
[1]Department of Computer Science and Engineering, SASTRA University, Srinivasa Ramanujan Centre, Kumbakonam-612 001, Tamilnadu, India
[2]B.Tech Computer Science and Engineering, SASTRA University, Srinivasa Ramanujan Centre, Kumbakonam-612 001, Tamilnadu, India

Abstract: Multiple Sequence Alignment (MSA) is an important part of bioinformatics domain in which two or more biological sequences, such as proteins, DNA or RNA are aligned sequentially. This Multiple Sequence Alignment plays a vital role in the generation of phylogenetic tree as well as predicting the protein structure. The protein sequences are generally used to generate the Phylogenetic tree. To attain this, we have transformed the protein sequences into numerical values using a substitution matrix and optimized those numerical values using Particle Swarm Optimization (PSO) method. The PSO is a meta-heuristic computational approach for performing optimization. The PSO uses the random values for pair-wise sequence alignment, resulting in decrease in the rate of the residues matched. This study presents how the rate of matching process can be improved by replacing the random values with the substitution matrix values if there is a positive value in the matrix. As result of this, we have also found that the score of the alignment sequence has been improved.

Keywords: BLOSUM50 substitution matrix, MSAPSO algorithm, multiple sequence alignment, particle swarm optimization, phylogenetic tree, protein sequence, velocity matrix

INTRODUCTION

Information about various biological sequences has grown exponentially in recent years (Das, 2008). The sudden rise of these biological data has increased the certainty of using a processing device and automation for storage. This has led to the door of the bioinformatics domain that offers a combination of biology and computers. One of the most familiar research areas of the bio-informatics domain is the sequence alignment. The sequence alignment generally refers to the alignment of protein/DNA molecules that results in the generation of either a phylogenetic tree or prediction of the protein structure. The main aim of generating the phylogenetic tree is that we can bring out the evolutionary relationship among the protein family. The proteins are generally the combination of amino acids. The sequence alignment is of two types such as Pair-wise Sequence Alignment (PSA), which deals with alignment of only two sequences and Multiple Sequence Alignment (MSA) (Hogeweg and Hesper, 1984) that deals with alignment of two or more sequences. The sequence alignment is generally meant for extracting meaningful information about two or more sequences that highlight their relationships by matching their residues that are similar and by inserting gaps if necessary. This sequence alignment can be attained either by dynamic programming or progressive methods or iterative methods or motif finding. Out of which, the iterative method is the best method for attaining an accurate sequence alignment when compared to progressive method. The main idea behind the study is to improve the alignment score by finding out similarities among two or more sequences.

THE SEQUENCE ALIGNMENT PROBLEM SPECIFICATION

During the process of evolution, insertion, deletion or mutation of the residues is possible. Thus, in order to highlight the similarities among sequences, it is often convenient to insert gaps in them, leading to a higher number of symbol matches. The similarity rate of aligned sequences is measured using a scoring function, which is based on a matrix that assigns a score to every pair of symbols. The process of finding an optimum match between the sequences is called sequence alignment. In recent years Computational Intelligence (CI) has provided an adaptable behavior for the systems within the changing environment. One of such approach that is proven to be successful is the Particle Swarm Optimization (PSO). With the help of PSO algorithm, the sequences are aligned by means of position vector and direction for optimization. There are many methods proposed by many people, for performing a multiple sequence alignment.

Corresponding Author: K. Arulmani, Department of Computer Science and Engineering, SASTRA University, Srinivasa Ramanujan Centre, Kumbakonam-612 001, Tamilnadu, India

One of such methods was proposed in Needleman and Wunsch algorithm (Needleman and Wunsch, 1970; Wang-Sheng and Shun-Feng, 2008), which is a standard algorithm for sequence alignment, where there are some drawbacks of time and space complexities. To overcome these problems, an MSAPSO algorithm was proposed. In this study, the main drawback was due to the use of random value generation even for matched residues, resulting in decrease in the rate of matched residues and in the score of the alignment. To improve the matching process, the random values are replaced by the values in the substitution matrix. As a result of this, the score of the alignment is also improved. A common heuristic is to seek a multiple alignment that maximizes the SP score (the summed alignment score of each sequence pair), which is NP complete. Therefore, the design of algorithms for multiple sequence alignment has been a very active research area.

IMPLEMENTATION OF REFINED MSAPSO

To attain MSA with good score, we go for numerical transformation of residues as the first step. Here, the residues are converted into numerical values with the help of predefined substitution matrix values. Then, according to the velocity vector and position of the residues, a new updated velocity matrix is created using the PSO technique. By comparing the values in the updated velocity matrix, the sequences are aligned. MSA is attained, by following the iterative principle for the given family of protein sequence. While obtaining the MSA, the residue match is done based on the refined MSAPSO, where the matching of residues has been improved, which in turn increases the score of the alignment.

Numerical transformation of residues: The protein sequence is the composition of amino acids. But for optimization techniques numerical values have to be given as input. So, the amino acids sequences are converted into numerical values. Generally, there are many predefined substitution matrix tables available to calculate these numerical values. We use BLOSUM50 substitution matrix, which are biologically and chemically tested. Based on this matrix, the first two residues are taken from the first two sequences and a second matrix is generated. Let the two sequences be S1 and S2 in which the length of S1 is X and length of S2 is Y. So, the numerical transformation takes place till X and Y. The obtained numerical values are termed as velocities and the matrix is termed as velocity matrix.

Generating updated velocity matrix using PSO: Particle Swarm Optimization (PSO) was first introduced by Kennedy and Eberhart (1995) and partly inspired by the behavior of large animal swarms such as schooling fish or flocking birds. PSO conducts search using a population of a random solutions, corresponding to individual. In addition, each potential solution called particles is also assigned a randomized velocity. Each particle in PSO flies in the hyperspace with a velocity, which is dynamically adjusted according to the flying experiences of its own and its colleagues.

Each particle adjusts its position iteratively according to their own or their neighboring particles towards two points: the best position of its own called the Pbest and by the best of all Pbest called Gbest. The particle swarm optimization concept consists of, at each time step, changing the velocity each particle toward its Pbest and Gbest. This algorithm is generally used to reduce the search space. The optimization technique is applied to the velocities in the velocity matrix. By changing its position and velocity vector the alignment of residues is done in optimization. Based on these positions and velocity vectors, a new velocity is calculated and a new updated velocity matrix is generated using PSO algorithm (Lakshmi Jagadamba *et al*., 2011; Wang-Sheng and Shun-Feng, 2008; Henikoff and Henikoff, 1992).

Implementation of refined MSAPSO: The proposed refined MSAPSO algorithm will use the advantages of MSAPSO such as faster alignment, less use of memory and also eliminating the problem of randomness. In this algorithm, we generate PSA for two sequences from a given protein sequence family and by repeating this process recursively MSA is attained by using refined MSAPSO. To implement our first step in refined MSAPSO method, we have to give two sequences S_1 and S_2 of length X and Y respectively as input. From this we get, an aligned sequence with insertion of gaps of same length l, where l>= max (X, Y). By following these steps we can generate PSA:

Consider two residues from each of the sequence (say i-1,i from S_1,j-1 and j from S_2) (Lakshmi Jagadamba *et al*., 2011):

1 Initialize

- Pbest P_b to zero
- Gbest G_b to X
- Position P = X+Y-2- (current position in the S_1 + current position in the S_2 + 2)

2 Now consider the first two residues of sequence S_1 and S_2, apply the numerical transformation and create a 2X2 matrix (Lakshmi Jagadamba *et al*., 2011).

3 Apply the PSO algorithm and generate updated velocity values (Lakshmi Jagadamba *et al*., 2011) and consider those values as current velocity values for the residues. Find the maximum value from it and follow these steps:

- Now check whether the index positions (i, j) and (i-1, j-1) of the second matrix value is positive, if so then

Table 1: Protein sequences that are used from uniprot database

Family name	Description	No. of sequences tested
NGF	Nerve growth factor family.	10
Globin	Family of proteins involved in binding and transporting proteins.	10
Myosin	Family of motor proteins.	9
G protein	Family of proteins for transmitting signals outside the cell.	8
HSP60	Family of heat shock proteins.	11
Dynein	Family of motor proteins for converting chemical energy to mechanical energy.	9
Kinesin	Family of motor proteins.	8

set the updated velocity value as maximum V_{max} and perform the alignment as follows:

- ◦ Align the residues i to j if the M (i, j) is maximum.
- ◦ Align the residues i to j and i-1 to j-1 if the M (i-1, j-1) is maximum.
- Else based on V_{max} perform the following steps:
- ◦ Align the residues i to j if the M (i, j) is maximum.
- ◦ Align the residues i to j and i-1 to j-1 if the M (i-1, j-1) is maximum.
- ◦ Insert a gap if in S_1 i.e., align gap to residue j if the M (i-1, j) is maximum.
- ◦ Insert a gap if in S_2 i.e., align gap to residue j if the M (i, j-1) is maximum.

4 Update

- pbest P_b = Round (mod (V_{max}))
- Gbest $G_b = G_b+g$

Repeat step 2 and 3 until it covers all the residues in the sequence.

In this algorithm, we have created a new technique of updating the current velocity value, which helps in increasing the residue match. This helps for increasing the score of even the closely related protein sequences. By giving this obtained PSA of various lengths X_1, X_2, X_3, X_4, ..X_n as input, we can get an aligned MSA by inserting gaps of length $k \geq = \max(X_1, X_2, X_3, X_4, ..X_n)$. Now, a pair-wise aligned sequences will be obtained. From the PSA, following steps creates the MSA:

5 Repeat the steps 1 to 4 iteratively, for all the protein sequences.

6 Now calculate the biological distance for all the protein sequence pair that is obtained from the above step. Store the value in an array of length ((X* (X-1))/2).

7 Then a distance matrix is created using the biological distance.

8 Using the distance matrix a phylogenetic tree is created using algorithms such as Neighbor-Joining.

9 Build an initial alignment with closely related sequences and gradually add the distance related sequences.

10 Repeat the process, until all the sequences are aligned, insert gap when necessary.

The above process will help us to generate a MSA with the help of the basic method Neighbor-joining. Instead of Neighbor-Joining method, we can use any other cluster method if desired.

Table 2: Comparison of score value between refined MSAPSO with MSAPSO

Family	No. of sequences tested	Score of match	
		Old	New
NGF 10	27.36	54.97	
Globin	10	41.07	68.00
Myosin	9	6.19	15.12
G protein	8	3.75	8.38
HSP60	11	9.56	18.73
Dynein	9	8.00	16.42
Kinesin	8	47.42	96.74

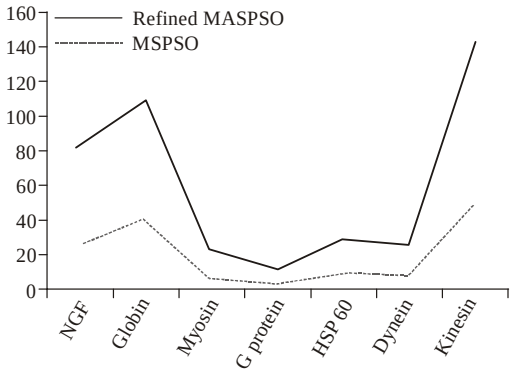

Fig. 1: Score of refined MSAPSO VS MSAPSO

EXPERIMENTAL ANALYSIS AND RESULTS

In this study, we have used only protein sequences as input. These families of protein sequences are extracted from uniprot database (http://www.uniprot.org/ last visited on 18.04.2012). From that database, we have tested only for few protein families as shown in the Table 1:

In order to check whether we got a better alignment or not, the alignment score is calculated (Lakshmi Jagadamba *et al*., 2011) by using the formula:

Score of the alignment = Number of residues matched/Total number of residues compared in the alignment.

After calculating the score value, when it compared with the normal MSAPSO algorithm, it has been found that there is a drastic increase in score rate for closely related sequences. The Table 2 shows the comparison of score value between MSAPSO and refined MSAPSO and the corresponding graph is shown in the Fig. 1.

It is obvious through the graph (Fig. 1) that the rate of residue matching by refined MSAPSO has improved when compared to normal MSAPSO.

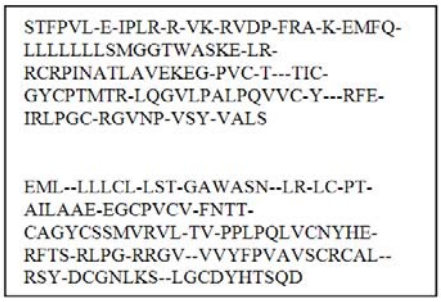

Fig. 2: Alignment using MSAPSO

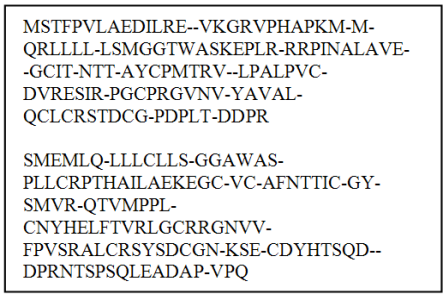

Fig. 3: Alignment using refined MSAPSO

The Fig. 2 and 3, shows the alignment of normal MSAPSO algorithm and refined MSAPSO algorithm respectively.

CONCLUSION

The refined MSAPSO algorithm has improved the score efficiently. Due to the improvement in the alignment score, we were able to get a better and an aligned MSA which would be very useful for generation of phylogenetic tree. We have also done some concepts of optimization too using the PSO algorithm.

REFERENCES

Das, S., 2008. Swarm Intelligence Algorithms in Bioinformatics. Studies in Computational Intelligence (SCI), 113-147.

Henikoff, S. and J.G. Henikoff, 1992. Amino acid substitution matrices from protein blocks. Proc. Atl, Acad. Sci. USA89, 22: 10915-10919.

Hogeweg, P. and B. Hesper, 1984. The alignment of sets of sequences and the construction of phyletic trees an integrated method. J. Mol., E20: 175-186.

Lakshmi Jagadamba, P.V.S., M.S.P. Babu, A.A. Rao, T.M.N. Vamsi and P.K.S. Rao, 2011. An improved algorithm for multiple sequence alignment using particle swarm optimization. Software Engineering and Service Science (ICSESS) IEEE 2nd International Conference, pp: 544-547.

Needleman, S.B. and C.D. Wunsch, 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Bio., 48: 443-453.

Wang-Sheng J. and S. Shun-Feng, 2008. Multiple sequence alignment using modified dynamic programming and particle swarm optimization. J. Chin. Inst. Eng., 31(4): 659-673.