A Novel Approach Based on Approximation and Heuristic Methods Using Multiple Sequence Alignments

KEYWORDS Multiple sequence alignment, Progressive alignment, Approximation technique.

Suresh.G
Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli-627 012, India

Vijayalakshmi.C
Department of Mathematics, VIT University, Chennai, Tamil Nadu.

ABSTRACT *Over the past decades, prominent methods were proposed for aligning biological sequences that finds the homology or similarity among them. One of the challenging task for solving problems in computational biology is the Multiple Sequence Alignment (MSA). In this paper, we discussed efficient methods designed to address the MSA problem. An algorithmic approach to MSA based on the combination of approximation and progressive techniques is compared. Particularly, the well-known algorithms such as Centre star Alignment and heuristic methods were used to align sequences. Consequently, how an optimization approach to MSA based on Particle Swarm Optimization (PSO) algorithm is discussed. Finally, using an empirical illustration, the above techniques were employed in order to improve the quality of alignment solutions. The results were shown numerically and graphically.*

1. Introduction

The primary goal of bioinformatics seeks to compare DNA / protein sequences, to search for a common or related pattern among them. During recent years, MSA methods in bioinformatics act as a challenging task for solving problems in computational biology for biological sequences such as biological macromolecules, DNA and proteins (Desmond, G. Higgins et al., 1988; Geoffrey, J. Barton, 1998; Julie, D. Thompson, et al; 1994; Wang-Sheng Juang and Shun-Feng Su. 2008). Thus, sequence alignment techniques are considered as central to modern molecular biology. The multiple alignment methods can be divided into two main categories: methods aligning sequences over their entire length (global) and methods aligning regions of only high similarity (local). In this paper, we focus in global alignment.

Needleman and Wunsch (1970) are often attributed as the first application of dynamic programming in molecular biology, while various formulations of the same algorithm were described by Sellers (1974) and Waterman et al., (1976). Desmond G. Higgins et al., (1988) performed multiple alignments of large numbers of amino acid or nucleotide sequences from a series of pairwise alignments of clusters, following the order of phylogeny tree. It is proved that the method is sufficiently fast and economical with memory on microcomputers. Julie D.Thompson et al., (1994) used progressive multiple sequence alignment method for the alignment of divergent protein sequences through sequence weighting, position specific gap penalty and weight matrix choice using CLUSTALW. Geoffrey J. Barton (1998) discussed the basic algorithms for alignment of two or more protein sequences and described alternative methods for scoring substitutions and gaps for local and global alignment methods. Wang-Sheng Juang eta l., (2008) presented a hybrid algorithm for multiple sequence alignment method and proposed an algorithm named Modified Dynamic Programming and Particle Swarm Optimization (MDPPSO).

In this paper, Multiple Sequence Alignment problem was incorporated effectively based on three well known techniques. The objective of this paper was fulfilled in three aspects. Initially, a MSA is carried out using approximation algorithm known as Centre-Star (CS) method, where to find the sequence that is the most similar to all the rest using pairwise alignments. Consequently, the well known progressive approach using Feng-Doolittle algorithm is implemented for the sequences, which are aligned progressively based on guide tree. Finally, the paper discussed the particle's position update based on iterative approach using Particle Swarm Optimization (PSO) algorithm. The theoretical and computational results were depicted in the respective sections.

2. Multiple Sequence Alignment

The MSA of DNA, RNA and protein sequences is one of the most common ad important tasks in bioinformatics. It is a process of aligning more than two sequences simultaneously, showing how the sequences are related to each other. MSA is also used for constructing evolutionary trees from DNA sequences and for analyzing the structures to help in designing new proteins. During recent decades, there has been increasing interest in the biosciences for methods that can efficiently solve this problem for sequences such as biological macromolecules, DNA and proteins. Recently, several programs are used to make multip1e sequence alignment automatically. Dynamic Prograrnming approach (DP), progressive method and statistical approach are commonly used in multiple sequence alignments. It is well known that the MSA problem can be exactly solved by the DP algorithm (Needleman and Wunsch; 1970), which converts the original problem to a problem of searching for the shortest path. But the drawback of the method is the excessive need for an increase in computation time and memory consumption in proportion to the increase in the number of sequences. This is where heuristic algorithms may take place in order to efficiently solve the sequence alignment problem.

There are several MSA algorithms reported in the literature (Chen, et al., 1992, McClure, et al., 1994, Thompson, et al., 1999). A great majority of MSA algorithms such as progressive, extension of DP, iterative and stochastic approaches (Simulated Annealing(SA), Genetic Algorithms(GA), Evolutionary Programming(EP)) are widely spread in bioinformatics research areas. Most of the approaches to MSA problem are based on the progressive approach proposed by Feng and Doolittle. This heuristic algorithm use pairwise alignments to construct a global alignment by aligning two more similar sequences, and then adding the other sequences one by one.

2.1 Definition of an MSA

The MSA, in general defined as notationally,

Let us consider s sequences $A_{i,} i=1,2\ldots S$ over an alphabet $\sum$. Then

$$A = \left| \begin{array}{l} A_1 = (a_{11}, a_{12}, \ldots, a_{1n}) \\ A_2 = (a_{21}, a_{22}, \ldots, a_{2n}) \\ \vdots \\ A_s = (a_{s1}, a_{s2}, \ldots, a_{sn}) \end{array} \right. \quad \ldots\ldots(1)$$

The MSA of A is obtained by inserting gaps (' – ') into original sequences such that all resulting sequences A_i^* have equal length $L \geq Max\{n_i \mid i = 1,2...S\}$, one can get back the sequence A$_i$ by removing all gaps from A_i^*, and no column consists of gaps only, then

2.2. Representation of MSA

$$A^* = \begin{cases} A_1^* = (a_{11}^*, a_{12}^*, \ldots a_{1L}^*) \\ A_2^* = (a_{21}^*, a_{22}^*, \ldots a_{2L}^*) \\ \vdots \\ A_S^* = (a_{S1}^*, a_{S2}^*, \ldots a_{SL}^*) \end{cases} \quad \ldots\ldots(2)$$

Assume that we want to align four sequences $(S_1, S_2, S_3 \; and \; S_4)$ which are as follows:

S_1	A	G	T	T	C	A	G		
S_2	A	T	G	T	C	A	G		
S_3	A	G	G	T	G	C	A	G	G
S_4	A	T	G	C	C	A	T		

Figure 1: An illustration of multiple sequences (DNA)

For the above multiple sequences, the alignments are made which includes insertion, deletion and gaps and that take into account the degree of variation in all sequences. Carrying out MSA mechanism is practically intractable as the number of sequences increases (which was discussed in Section 2). Now, the global alignment (methods aligning sequences over their entire length) of the considered sequences are,

-	-	A	T	G	T	C	A	G
-	-	A	T	G	C	C	A	T
-	A	G	T	T	C	A	G	-
A	G	G	T	G	C	A	G	G

Figure 2: Global alignment of sequences

2.3. Sum-of-Pairs (SP) Scores

The standard method for scoring alignments is not Hidden Markov Model (HMM) formulations, but is similar in that it does not use a phylogenetic tree and it assures statistical independence for the columns. Consider two sequences A_p^* and A_q^* in the alignment. For two aligned symbols u and v, then we define,

$$S(u,v) = \begin{cases} match \;\; score \;\; for \; u \; and \; v, if \; u \; and \; v \; are \; residues \\ -d, if \; either \;\; u \; or \; v \; is \; a \; gap, or \\ 0, if \; both \;\; u \; and \; v \; are \; gaps. \end{cases} \quad \ldots\ldots(3)$$

It is to be noted that (u = - and v = -) can occur simultaneously in a multiple alignment. For pairwise alignment of sequences A_p and A_q imposed by a multiple alignment A^* of r sequences, then denote the score of this induced pairwise alignment as,

$$S(A_p^*, A_q^*) = \sum_{i=1}^{L} S(a_{pi}^*, a_{qi}^*) \quad \ldots\ldots(4)$$

In general, the sum-of-pairs (SP) scores of an alignment is defined by,

$$S(A_1^* \; \ldots\ldots A_r^*) = \sum_{1 \leq p \leq q \leq r} S(A_p^*, A_q^*) \; \ldots\ldots(5)$$

3. Approximation Approach

Among the various methods like global optimization, approximation, heuristic and probabilistic methods for solving the MSA problem, the approximation technique is known for its simplicity and straightforward. In this section, how the approximation algorithm was implemented using Centre Star (CS) method will be discussed as follows.

The primal idea of star alignment method is to find the sequence that is the most similar to all the rest using pairwise alignment and use it as the 'centre of a star' when aligning all other sequences. The CS method works in two steps: (i) Identify the centre string S$_c$ of S and (ii) Uses the alignments of S$_c$ with each S$_i$ to create a multiple alignment. The Centre Star algorithm includes the following steps:

(i) To find $D(s_i, s_j) \; \forall i, j.$

(ii) To find the centre sequence S$_c$ which minimizes $\sum_{i=1}^{k} D(S_c, S_i)$

(iii) For every $S_i \in S - \{S_c\}$, choose an optimal alignment between S$_c$ and S$_i$.

(iv) Introduce spaces into S$_c$ so that the multiple alignment M satisfies the alignments found in step3.

It is noted that for consistency, once a gap is added, it is never removed and the running time is dominated by computing the pairwise alignments. If k sequences are length n, then compute $\frac{k(k-1)}{2}$ pairwise alignments and each alignment consumes time n2. Thus, the running time for computing all pairwise alignments is $O(k^2 n^2)$. Also, it is noted that the Star Alignment does not optimize the Sum of Pairs Scores criterion and therefore, the method is not much of use in practical sense. The similar method that is very widely used is the progressive alignment method of the Clustal W package. However, it compares pairwise alignments and then adds individual sequences based on a given order of similarity. In the next section, the description of this method is presented.

4. Progressive Approach

Multiple alignments act as a key factor for the prediction of protein secondary structure, residue accessibility, function and the identification of residues important for specificity. It also provides the basis for the most sensitive sequence searching algorithms (Barton, G.J and Sternberg, M.J.E, 1990; Gribskov, M. et al., 1987). From theoretical point of view, it is well-known that MSA problem can be exactly solved by the dynamic programming algorithm (Needleman and Wunsch, 1970; Smith and Waterman, 1981) for local and global alignments which ensures an optimal alignment of the sequences. In practice, this technique leads a trivial task for aligning more than three or four sequences because the computational cost increases exponentially with the number of dimensions.

Therefore, most of the approaches to MSA problem are based on the progressive approach proposed by Feng and Doolittle (1987). The basic idea behind the method is that it does not compare all of the sequences together, rather use pairwise alignments to constructs a global alignment by aligning the two more similar sequences, and then adding the other sequences one by one. The most popular program used for MSA is ClustalW, the computational procedure is described in the form of flowchart, shown in Figure 2.

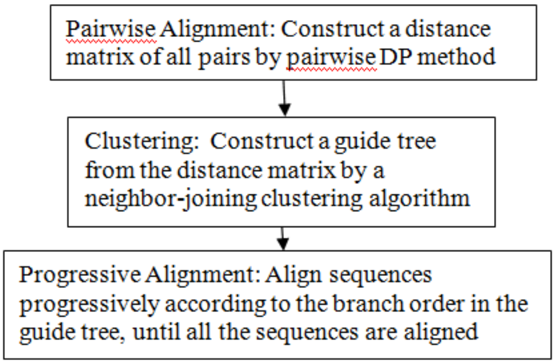

Pairwise Alignment: Construct a distance matrix of all pairs by pairwise DP method
Clustering: Construct a guide tree from the distance matrix by a neighbor-joining clustering algorithm
Progressive Alignment: Align sequences progressively according to the branch order in the guide tree, until all the sequences are aligned

Figure 3: The Basic Progressive Alignment Procedure

5. Iterative Approach

In this section, an algorithmic approach based on iterative technique for MSA is discussed. In computational intelligence, some algorithms based on the social behavior of animals, such as ants, bees, birds and human beings have been developed [9]. Among them, one of the well-known algorithms is called Particle Swarm Optimization (PSO). The theory was first proposed by James Kennedy and Russel Eberhart [5] in 1995 and is motivated as a model of the social behavior of organisms such as bird flocking and fish schooling. Based on such an idea, many researchers modified and applied the theory to widespread areas (Abido, 2002, Kumar, et al., 2004, Robinson and Rahmat-Samii, 2004, Shi and Eberhart, 1998, Van der Merwe and Engelbrecht, 2003). PSO is a population based heuristic search technique in which each particle represents a potential solution within the search space and will be characterized by its position, its velocity and a record of its past performance. The PSO algorithm consists of three steps, which are repeated until the termination criterion is met.

1. Evaluate the fitness of each particle
2. Update individual and global best fitness and positions
3. Update velocity and position of each particle.

A general representation of the basic PSO algorithm proposed (Kennedy and Eberhart, 1995) is described as below,

x_k^i - particle position

v_k^i – particle velocity

p_k^i – best remembered individual particle position

p_k^g – best remembered swarm position

c_1c_2 – cognitive and social parameters

r_1r_2 – random numbers between 0 and 1

The position of individual particles updated as follows,

$$x_{k+1}{}^{i} = x_k{}^{i} + v_{k+1}{}^{i} \qquad ....(6)$$

with the velocity calculated as,

$$v_{k+1}{}^{i} = v_k{}^{i} + c_1r_1(p_k{}^{i} - x_k i) + c_2r_2(p_k{}^{g} - x_k{}^{i}) \quad ....(7)$$

To implement PSO for MSA [11], each particle in the problem space represents a string of gap positions $X = x_1^1x_2^1...x_{n1}^1$, $x_1^2x_2^2...x_{n2}^2...x_1^mx_2^m...x_{nm}^m$, where x_j^i for $1\leq j\leq n_i$ and $1\leq i\leq m$ is the location of a gap existing in sequence i. Here, m is the number of sequences and n_i is the number of gaps for sequence i. And n_i is obtained as $L – l_i$, where l_i is the length of the ith original sequence and L is the length of the sequences used in the algorithm. Finally, the PSO algorithm maintains the best fitness value achieved among all particles in the swarm, called the global best fitness, and the candidate solution that achieved this fitness, called the global best position or global best candidate solution.

6. Empirical Illustration

In this section, an algorithmic approach to MSA is carried out using the combination of approximation, heuristic and optimization methods. Initially, the centre star method of approximation algorithm is carried out for the multiple sequences to find the sequence that is the most similar to the rest. Consequently, the heuristic algorithm based on the progressive approach (proposed by Feng and Doolittle, 1987) is performed for aligning the same set of sequences and sum-of-pairs scores were computed. Finally, the PSO technique will be discussed to improve the alignment of sequences based on the modified particles swarm optimizer (proposed by Kennedy and Eberhart, 1995).

To illustrate the way of employing approximation algorithm to MSA, the set of sequences depicted in Figure 1 is considered again. Initially, the pairwise alignments scores for all pairs is computed using the simple scoring scheme (match = 1, mismatch = -1 and gaps = -2), which is shown in Table 3. Here, S_1 is the sequence most similar to the rest, and below are the best alignments between S_1 and the rest of the sequences. Having identified the centre string S_c of S, the alignments of S_c with each S_i is created for multiple alignments. The result is shown in Figure 4. Based on the algorithm steps depicted in section 3, the rest of the computational results are obtained and shown in Figures 5-7.

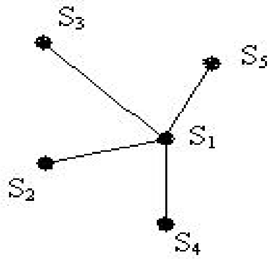


Figure 4: S_1 is the centre of

	S1	S2	S3	S4
S1	0	3	-1	-1
S2		0	-5	3
S3			0	-5
S4				0

Figure 5: Shows the pairwise alignment scores

Sequences to find S_c such that between every pair of sequences

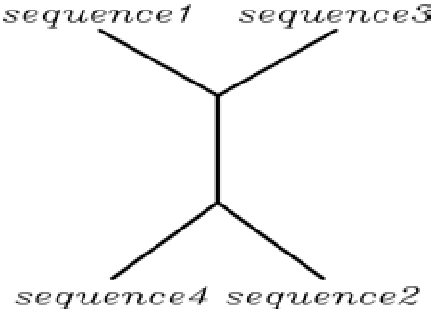
$\sum_{i\neq c}(s_i, s_c)$ is maximized.

S1 S2	A G T T C A G A T G T C A G
S1 S2 S3	- A G T T C A G - A G G T G C A G G - A T G T C A G -
S1	- - A T G T C A G
S2	- - A T G C C A T
S3	- A G T T C A G -
S4	A G G T G C A G G

Figure 6: Pairwise alignments with the Figure 7: Shows the Multiple alignments of all

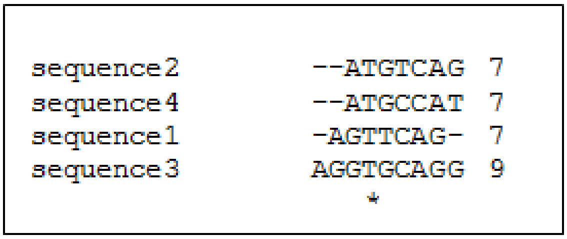
Centre sequence S1, by adding sequences which is consistent with S2 and S3 all pairwise alignments

Based on ClustalW (the most popular tool used for MSA), the following results are obtained using the same set of sequences followed by the algorithm procedure presented in section 4. Firstly, the pairwise similarity scores are constructed and based on the scores, the guide tree was obtained from the distance matrix by NJ method [3]. Then sequences are aligned progressively until all the sequences are aligned according to the guide tree constructed. The results based on ClustalW are shown below



Sequences		Alignment score
1	2	57.142
1	3	71.428
1	4	28.571
2	3	42.857
2	4	71.428
3	4	42.857

Figure 8: Pairwise distance scores of sequences Figure 9: Guide tree based on NJ method



```
sequence2      --ATGTCAG  7
sequence4      --ATGCCAT  7
sequence1      -AGTTCAG-  7
sequence3      AGGTGCAGG  9
                  *
```

Figure 10: Results of multiple sequence alignments using ClustalW.

Single asterisks * indicates the highlight sequence similarities among the residues.

Finally, the paper proposed to use pairwise centre-star alignments and to incorporate PSO into the algorithm. Using the centre star method of multiple alignments of all sequences (shown in Figure 7), it is found that the length of the sequence L is 9. According to Figure 7, sequences S1and S2 have 2 gaps in each positions 1 and 2; S3 have 2 gaps in positions 1 and 9 and whereas S4 possess no gaps in any positions. According to this gap position information, a particle can be coded into an integer number vector [1, 2, 1, 2, 1, 9, 0]. Here, the length of sequence is 9 and each one of the other N-1 particles will be formed like [$a_{1,}$ a_2, b_1, b_2, c_1 $c_{2,}$ d] are randomly generated within the integer range [1, 9]. It is noted that the particle positions thereby can be updated by Eqn.1. According to Figure 7, after four sequences are all included, then we have $x_{id}(0)$, L, $l_{1,}$ l_2, l_3, n_1, n_2, n_3 as [1, 2, 1, 2, 1, 9, 0], 9, 7,7,7,9,2,2,2,0. In a similar way, each particle's position is updated by applying the new velocity to the particle's previous position based on Eqn.1. After the update of particles, the sum-of-pairs scores (shown in section 2.3) of the final global sequence alignment is computed as 40.

Results and Discussions

In this section, the results obtained in the previous illustration were discussed. The primitive objective of the paper was fulfilled in three aspects. The description is as follows.

(i) One of the well-known approximation algorithms using Centre-Star method is applied to the multiple sequences. Based on the computational steps discussed in section 3, it is found that sequence S1 is similar to the rest of the other three sequences and hence it is considered as 'centre of a star' (shown in Figure 4). After computing the pairwise similarity scores using the scoring scheme strategy, alignments are made. Initially, S1 and S2 are aligned, then S3 is added using its alignment to S1 as (S1, S2, S3). Then, the final sequence S4 is added using its alignment to S1 as (S1, S2, S3, S4). Hence, the multiple alignments of all sequences obtained is shown in Figure 7.

(ii) Most packages use heuristics to compute multiple sequence alignments. One of the commonly used progressive approach by Feng-Doolittle algorithm is applied to multiple sequences. For computational purpose, ClustalW is used to align sequences progressively using three steps described in section 4. After computing the half distance matrix of n(n-1) distances, a guide tree is constructed. From Figures 8 and 9, it is found that sequences S1;S3 and S2;S4 possess high similarity score as 71.42% when compared to other pairs of sequences. According to the branch order in the guide tree, the sequences are aligned progressively until all sequences are aligned, which is shown in Figure 10 (Results obtained from ClustalW).

(iii) The final goal is employed with the implementation of pairwise centre star alignment method and to incorporate particle's positions updation using PSO algorithm. In the basic particle swarm optimization algorithm, particle swarm consists of "n" particles, and the position of each particle stands for the potential solution in D-dimensional space. For this illustration, the PSO technique is adopted after the multiple alignments of all sequences using centre star method (shown in Figure 7), it is found that sequence S4 have no gaps whereas S1, S2, S3 possess two gaps each in different positions. After the particles position is updated by applying the new velocity to the particle's previous position based on Eqn.1, the sum-of-pairs scores of the final global sequence alignment is computed.

Conclusions

In this paper, three techniques such as approximation, heuristic and optimization were incorporated to address the MSA problem. The theoretical discussions and computational results obtained using the above techniques were presented in their respective sections. When approximation approach using centre star alignment algorithm is compared to heuristic approach using Feng-Doolittle algorithm, it is evident that the two methods were similar. Using centre-star method, the running time for all pairwise alignments is $O(k^2n^2)$ and it does not optimize sum-of-pairs scores criterion. Thus, the method is practically not used much. ClustalW uses the same basic method as the Star Alignment. However, it compares pairwise alignments and then adds individual sequences based on a given order of similarity.

When focusing about PSO, it is a heuristic global optimization method and also an optimization algorithm, which is based on swarm intelligence. The search can be carried out by the speed of the particle. During the development of several generations, only the most optimist particle can transmit information onto the other particles, and the speed of the researching is very fast. It is concluded that, when compared with the other standard dynamic programming methods to MSA, the progressive/ hierarchical approach and iterative approach have a greater advantage because of its simplicity, speed and flexibility.

REFERENCE

[1] Bains, W.(1986). 'MULTAN: A program to align multiple DNA sequences'. Nucleic Acids Research. Vol.14. pp.159-177. | [2] Desmond, G.Higgins and Paul M, Sharp. (1988). 'CLUSTAL: A package for performing multiple sequence alignment on a microcomputer'. Gene. pp. 237-244. | [3] Feng, D.F. and Doolittle, R.F. (1987). 'Progressive sequence alignment as a prerequisite to correct phylogenetic trees'. Journal of Molecular Evolution. Vol.25. pp. 351-360. | [4] Geoffrey J. Barton. (1998). 'Protein sequence Alignment Techniques'. Acta Cryst. Vol.54. pp. 1139-1146. | [5] James Kennedy and Russell Eberhart. (1995). ‘Particle swarm optimization’. In Proceedings of the IEEE International Conference on Neural Networks, Vol.4. pp. 1942–1948. | [6] Julie, D. Thompson, Desmond G. Higgins and Toby J. Gibson (1994). 'CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice'. Nucleic Acids Research. Vol.22. pp. 4673-4680. | [7] Lipman, D.J., Altschul, S.F. and Kececioglu, J.D.(1989). 'A tool for multiple sequence alignment'. Proceedings of the National Academy of Sciences of the United states of America. Vol.86. pp. 4412-4415. | [8] Needleman, S.B. and Wunsch, C.D.(1970). 'A general method applicable to the search for similarities in the amino acid sequence of two proteins'. .Journal of MolecularBiology. Vol.8. pp. 443-453. | [9] Pedro F. Rodriguez, Luis F.Nino and Oscar M.Alonso.(2007). ‘Multiple sequence alignment using swarm intelligence’. International Journal of computational intelligence. Vol.3.pp. 123-130. | [10] Qinghai Bai.(2010). ‘Analysis of Particle Swarm Optimization Algorithm’. Computer and Information science. Vol.3. pp. 180-184. | [11] Wang-Sheng Juang and Shun-Feng Su.(2008). 'Multiple sequence alignment using modified dynamic programing and particle swarm optimization'. Journal of the chinese institute of Engineers. Vol.31. pp. 659-673.