CHAPTER IV

MULTIPLE SEQUENCE ALIGNMENT USING APPROXIMATION AND HEURISTIC TECHNIQUES

4.1 INTRODUCTION

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) 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, the multiple sequence alignment of DNA, RNA and protein sequences play a significant role in bioinformatics is to develop software tools to generate useful biological knowledge.

Multiple Sequence Alignment 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 multiple sequence alignment automatically. Dynamic Programming 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 multiple sequence alignment algorithms reported in the literature [22] [32] [50] [89]. A great majority of MSA algorithms such as Progressive, extension of DP, Iterative and Stochastic approaches (Simulated Annealing (SA), Genetic Algorithms (GA), and 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 beginning with the most similar pair and progressing to the most distantly related, which finally builds up a MSA solution. When compared with the other standard dynamic programming methods to MSA, the progressive or hierarchical approach have a greater advantage because of its simplicity, speed and flexibility.
4.2 MULTIPLE SEQUENCE ALIGNMENT (MSA)

The MSA, in general defined as follows,

Let us consider \( s \) sequences \( A_i, i=1,2...S \) over an alphabet \( \Sigma \). Then

\[
A = \begin{cases}
A_1 = (a_{11}, a_{12}, ... a_{1n}) \\
A_2 = (a_{21}, a_{22}, ... a_{2n}) \\
\vdots \\
A_s = (a_{s1}, a_{s2}, ... a_{sn})
\end{cases} \tag{4.1}
\]

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

\[
A^* = \begin{cases}
A_1^* = (a_{11}^*, a_{12}^*, ... a_{1L}^*) \\
A_2^* = (a_{21}^*, a_{22}^*, ... a_{2L}^*) \\
\vdots \\
A_s^* = (a_{s1}^*, a_{s2}^*, ... a_{sL}^*)
\end{cases} \tag{4.2}
\]

Let us assume the alignment of four sequences,

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_1</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>S_2</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>S_3</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>S_4</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

Initially, the alignments are made by insertions, deletions 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
(discussed in Section 4.1). Now, the global alignment of the above sequences is shown in Figure 4.1.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>-</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
</tr>
</tbody>
</table>

Figure 4.1: Global Alignment of Multiple DNA sequences

4.3 ILLUSTRATION USING CLUSTALW

Data source

This illustration uses UniProt, one such a central repository of protein data created by combining the Swiss-Prot, TrEmbl and PIR-PSD databases, which is accessible at http://www/uniprot/org. The five insulin protein sequences for different organisms with similar gene names and similar sequence length was chosen to illustrate the performance of CLUSTALW (the homepage displayed in Figure 4.4). The searched results are displayed in FASTA format shown in Table 4.1.

The sequences displayed in Table 4.1 are aligned and the results are shown in Figure 4.2 (code names of the sequences are given at the extreme left, whereas the right part indicates the respective protein sequences with varying colored alphabets, a single asterisks * indicates the highlight sequence similarities among the residues).
Table 4.1 Illustration of FASTA sequences derived from UniProt

<table>
<thead>
<tr>
<th>Accession</th>
<th>Species</th>
<th>Gene Symbol</th>
<th>Peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>sp</td>
<td>P01308</td>
<td>INS_HUMAN</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>MALWMRLPLLALLALWGDPAAAFAVNHLCGSHLVEALYVCGERGFFYTPKTRREAE DLQVGVQVELG</td>
<td>MALWMRLPLLALLALWGDPAAAFAVNHLCGSHLVEALYVCGERGFFYTPKTRREAE DLQVGVQVELG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P01321</td>
<td>INS_CANFA</td>
<td>Canis familiaris</td>
</tr>
<tr>
<td>MALWMRLPLLALLALWAPTRAFVNQHLCGSHLVEALYVCGERGFFYTPKARREVEDLQVDRVELA GAPEGGEQPLALEGA</td>
<td>MALWMRLPLLALLALWAPTRAFVNQHLCGSHLVEALYVCGERGFFYTPKARREVEDLQVDRVELA GAPEGGEQPLALEGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P01311</td>
<td>INS_RABIT</td>
<td>Oryctolagus cuniculus</td>
</tr>
<tr>
<td>MASLAALLPLALLVLCLDPQAFAVNFQHLCGSHLVEALYVCGERGFFYTPKSREVEELQVQAELGG GPAGGLOPSALELALQKRGI</td>
<td>MASLAALLPLALLVLCLDPQAFAVNFQHLCGSHLVEALYVCGERGFFYTPKSREVEELQVQAELGG GPAGGLOPSALELALQKRGI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P01322</td>
<td>INS1_RAT</td>
<td>Rattus norvegicus</td>
</tr>
<tr>
<td>MASLAALLPLALLVLCLDPQAFAVNFQHLCGSHLVEALYVCGERGFFYTPKSREVEELQVQAELGG GPAGGLOPSALELALQKRGI</td>
<td>MASLAALLPLALLVLCLDPQAFAVNFQHLCGSHLVEALYVCGERGFFYTPKSREVEELQVQAELGG GPAGGLOPSALELALQKRGI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sp</td>
<td>P01329</td>
<td>INS_CAVPO</td>
<td>Cavia porcellus</td>
</tr>
<tr>
<td>MALWMHLLTLALLALWGPNHTQAFVSRLGCNVLVEYSRCCDDFPPYDKRRELEDQPVEQTELG MLGAGGLQPLALEMALQKRGIVDQCCTGTCRHLQSYCN</td>
<td>MALWMHLLTLALLALWGPNHTQAFVSRLGCNVLVEYSRCCDDFPPYDKRRELEDQPVEQTELG MLGAGGLQPLALEMALQKRGIVDQCCTGTCRHLQSYCN</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.2: Results of Multiple Sequence Alignments Using ClustalW
One of the popular techniques to MSA known as Feng-Doolittle algorithm was implemented for the protein sequences shown in Table 4.1. The computational analysis is carried out using step-by-step procedure. Based on the similarity scores, CLUSTALW produced a phylogeny tree (shown in Figure 4.3) used to infer biologically relevant phylogenetic relationships between the sequences. At a final stage, the nearest two sequences, as indicated by the guide tree are aligned using a pairwise alignment method. For aligning those sequences, a standard dynamic programming algorithm called Needleman-Wunsch was implemented.

From the overall similarity scores between aligned pair of sequences, it is evident that first two sequences (sequence 1 - human and sequence 2 - canfa) possessed a high similarity score of 520.00 when compared with the other pairs of sequences. The respective results obtained using ClustalW are shown in Table 4.2 and Figure 4.3.

![Figure 4.3: Output of Phylogenetic Tree based on UPGMA produced by ClustalW](image)
Table 4.2: Results of Pairwise similarity scores of sequences using ClustalW

<table>
<thead>
<tr>
<th>SeqA</th>
<th>Name</th>
<th>Length</th>
<th>SeqB</th>
<th>Name</th>
<th>Length</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sp</td>
<td>P01308</td>
<td>INS_HUMAN</td>
<td>110</td>
<td>2</td>
<td>sp</td>
</tr>
<tr>
<td>1</td>
<td>sp</td>
<td>P01308</td>
<td>INS_HUMAN</td>
<td>110</td>
<td>3</td>
<td>sp</td>
</tr>
<tr>
<td>1</td>
<td>sp</td>
<td>P01308</td>
<td>INS_HUMAN</td>
<td>110</td>
<td>4</td>
<td>sp</td>
</tr>
<tr>
<td>1</td>
<td>sp</td>
<td>P01308</td>
<td>INS_HUMAN</td>
<td>110</td>
<td>5</td>
<td>sp</td>
</tr>
<tr>
<td>2</td>
<td>sp</td>
<td>P01321</td>
<td>INS_CANFA</td>
<td>110</td>
<td>3</td>
<td>sp</td>
</tr>
<tr>
<td>2</td>
<td>sp</td>
<td>P01321</td>
<td>INS_CANFA</td>
<td>110</td>
<td>4</td>
<td>sp</td>
</tr>
<tr>
<td>2</td>
<td>sp</td>
<td>P01321</td>
<td>INS_CANFA</td>
<td>110</td>
<td>5</td>
<td>sp</td>
</tr>
<tr>
<td>3</td>
<td>sp</td>
<td>P01311</td>
<td>INS_RABIT</td>
<td>110</td>
<td>4</td>
<td>sp</td>
</tr>
<tr>
<td>3</td>
<td>sp</td>
<td>P01311</td>
<td>INS_RABIT</td>
<td>110</td>
<td>5</td>
<td>sp</td>
</tr>
<tr>
<td>4</td>
<td>sp</td>
<td>P01322</td>
<td>INS1_RAT</td>
<td>110</td>
<td>5</td>
<td>sp</td>
</tr>
</tbody>
</table>
### Multiple Sequence Alignment by CLUSTALW

**General Setting Parameters:**

- **Output Format:** CLUSTAL
- Pairwise Alignment: FAST/APPROXIMATE, SLOW/ACCURATE

**Enter your sequences (with labels) below (copy & paste):** PROTEIN, DNA

Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

**Or give the file name containing your query**

---

### More Detail Parameters...

**Pairwise Alignment Parameters:**

**For FAST/APPROXIMATE:**

- K-tuple(word) size: 1
- Window size: 5
- Gap Penalty: 3
- Number of Top Diagonals: 5
- Scoring Method: PERCENT

**For SLOW/ACCURATE:**
Among 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.

Figure 4.4: The CLUSTALW Homepage

4.4 APPROXIMATION APPROACH USING MSA
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 Centre Star method works in two steps:

- Identify the centre string $S_c$ of $S$ and
- Uses the alignments of $S_c$ with each $S_i$ to create a multiple alignment.

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 $\binom{k(k-1)}{2}$ pairwise alignments and each alignment consumes time $n^2$. Thus, the running time for computing all pairwise alignments is $O(k^2n^2)$. 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 ClustalW 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.

### 4.5 PROGRESSIVE/HEURISTIC APPROACH USING MSA

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 computational procedure of this algorithm is displayed below.

| **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 4.5: Schematic Representation of Progressive Approach to MSA**
4.6 ITERATIVE APPROACH USING MSA

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 [70]. 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 in 1995 and is motivated as a model of the social behavior of organisms such as bird flocking and fish schooling.

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 Particle Swarm Optimization algorithm proposed (Kennedy and Eberhart, 1995) is described as below.

The position of individual particles updated is defined by,

\[ \mathbf{x}_{k+1}^i = \mathbf{x}_k^i + \mathbf{v}_{k+1}^i \]  \hspace{1cm} \ldots (4.3)

with the velocity calculated as,

\[ v_{k+1}^i = v_k^i + c_1 r_1 (p_k^i - x_k^i) + c_2 r_2 (p_g^i - x_k^i) \]  \hspace{1cm} \ldots (4.4)
Where,

\[ x_{ki} \] - particle position

\[ v_{ki} \] - particle velocity

\[ p_{ki} \] – best remembered individual particle position

\[ p_{kg} \] – best remembered swarm position

\[ c_1, c_2 \] – cognitive and social parameters

\[ r_1, r_2 \] – random numbers between 0 and 1

To implement PSO for MSA [89], each particle in the problem space represents a string of gap positions \( X = x_1^1x_2^1... x_{n_1}^1, x_1^2x_2^2... x_{n_2}^2... x_1^mx_2^m... x_{nm}^m \), where \( x_{ji} \) 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 \( i^{th} \) 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.

4.7 NUMERICAL 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 is considered.

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0</td>
<td>3</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>S2</td>
<td>0</td>
<td>-5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>0</td>
<td>-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td></td>
<td></td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Initially, the pairwise alignments scores for all pairs is computed using the simple scoring scheme \(\text{match} = 1, \text{mismatch} = -1\) and \(\text{gaps} = -2\), shown in Table 4.3.

Table 4.3: Pairwise alignment scores of Multiple sequences
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 4.6. Based on the algorithm steps depicted in section 4.4, the rest of the computational results are obtained and shown in Figures 4.7 and 4.8.

![Figure 4.6: Output of sequence alignment using Centre-Star method](image)

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>A G T T C A G</td>
<td>A T G T C A G</td>
<td>- A G T T C A G -</td>
</tr>
<tr>
<td>$S_3$</td>
<td>- A T G T C A G -</td>
<td>A G T T C A G -</td>
<td>A G T T C A G -</td>
</tr>
</tbody>
</table>

![Figure 4.7: Results of Pairwise alignments using Centre-Star method](image)

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>- - A T G T C A G</td>
<td>- - A T G C C A T</td>
<td>- A G T T C A G -</td>
<td>A G G T G C A G G</td>
</tr>
<tr>
<td>$S_2$</td>
<td>- - A T G C C A T</td>
<td>- - A T G C C A T</td>
<td>- - A T G C C A T</td>
<td>- - A T G C C A T</td>
</tr>
<tr>
<td>$S_3$</td>
<td>- A G T T C A G -</td>
<td>- A G T T C A G -</td>
<td>- A G T T C A G -</td>
<td>- A G T T C A G -</td>
</tr>
</tbody>
</table>

![Figure 4.8: Results of Multiple Alignments using Centre-Star method](image)
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 4.5. Firstly, the pairwise similarity scores are constructed and based on the scores, the guide tree was obtained from the distance matrix by NJ method. Then sequences are aligned progressively until all the sequences are aligned according to the guide tree constructed. The results based on ClustalW are shown in following Table 4.4, Figure 4.9 and Figure 4.10.

**Table 4.4: Results of Pairwise Distance scores of Multiple sequences**

<table>
<thead>
<tr>
<th>Sequences</th>
<th>Alignment scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

**Figure 4.9: Guide Tree based on NJ method for Multiple sequences**
4.8 COMPUTATIONAL RESULTS AND DISCUSSIONS

In this section, Particle Swarm Optimization (PSO) algorithm is incorporated and using centre star method of multiple alignments (shown in Figure 4.8) and it is found that the length of the sequence L is 9. Accordingly, sequences S1 and 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.4.3.

According to Figure 4.8, after four sequences are all included, then we have \(x_{ld}(0), L, l_1, l_2, l_3, n_1, n_2, n_3\) 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.4.4. After the updation of

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Alignment</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence 2</td>
<td>- - ATGTCAG</td>
<td>7</td>
</tr>
<tr>
<td>Sequence 4</td>
<td>- - ATGCCAT</td>
<td>7</td>
</tr>
<tr>
<td>Sequence 1</td>
<td>- AGTTCAG-</td>
<td>7</td>
</tr>
<tr>
<td>Sequence 3</td>
<td>AGGTGCAGG</td>
<td>9</td>
</tr>
</tbody>
</table>

*
particles, the sum-of-pairs scores of the final global sequence alignment is computed as 40.

In this section, the results obtained in the previous illustration were discussed. The primitive objective 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 4.4, 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 4.6). 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 4.8.

(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 4.5. After computing the half distance matrix of n(n-1) distances, a guide tree is constructed. From Figure 4.9, 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, shown in Figure 4.10 (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 4.6), 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.4.3, the sum-of-pairs scores of the final global sequence alignment is computed.

4.9 CONCLUSION

In this section, 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 Particle Swarm Optimization technique, 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.