Implementing Protein Sequence Alignment using PAM-250 Matrices

Rajbir Singh
Associate Prof. & Head
Department of IT
LLRIET, Moga

Sukhleen Kaur Bhasin
Student of M.Tech
Department of CSE
LLRIET, Moga

DheerajPal Kaur
Assistant Professor
Department of ECE
LLRIET, Moga

ABSTRACT
Multiple Protein Sequence is one of the most important problems in modern computational biology. The emphasis here is on the use of computers because most of the tasks involved in genomic data analysis are highly repetitive or mathematically complex. One of the largest areas of Bioinformatics and Data mining has been in the Protein Domain. These efforts have included protein Structure prediction, folding Pathway prediction, Sequence alignment, Substructure Detection and many others. Data storage became easier as the accessibility of large amount of computing power at low cost. The research in bioinformatics has accumulated large amount of data. As the hardware technology advancing, the cost of storing is decreasing. The biological data is available in different formats and is comparatively more complex.

In the present work, data mining solution is provided for the problem of protein sequence alignment. Different formats of sequences are studied and plain text format is chosen for the problem under consideration. Clustering methods are based on expressing similarity or dissimilarity of such sequences. The similarity of two protein sequences can be assessed by score of the best alignment of the sequences.

Scoring matrix accesses the replacement of one amino acid by another, accepted by natural selection. The replacement can be due the result of two distinct processes: i) occurrence of mutation in the portion of the gene template producing one amino acid of a protein. ii) acceptance of the mutation by the species (similar function). PAM (Accepted Point Mutations) is the scoring matrice that is used for the different computations. PAM-250 matrix is used for the problem under consideration. The matrix is frequently used to score aligned peptide sequences to determine the similarity of those sequences. The numbers given above were derived from comparing aligned sequences of proteins with known homology and determining the "accepted point mutations" (PAM) observed. Global and Local alignments are predicted along with the alignment score.

1. INTRODUCTION
The proteins made from the pairing of the nucleotides present in the living organisms helps in finding the genetic information between the two or more different organisms. This genetic information helps to gather the information about the common ancestors. Sequence alignment is a way which arranges the one sequence to the another to identify the similar regions in order to define the similarity between the sequences. Matrix is formed between the two sequences in order to attain the successive similarities. While alignment the causes of (dis)similarity is also attained between the sequences these mismatches are of type's mutations, gaps, insertion and deletions. The two different types of aligning the sequences is global and local alignment. In global alignment two sequences are aligned from end to end, whereas; in local alignment the region of the sequence with highest number of similarity is found. Needleman and Wunsch algorithm given in 1970 is used to attain global alignment; it is identified by drawing a matrix along x- and y- axis. For Local alignment Smith waterman algorithm is used which was discovered in 1981. The scoring system helps in deciding the good alignment, as the score of each alignment is calculated and the alignment with the highest score is considered as the best alignment. To attain more accurate sequences we use some substituted matrices values. As here in our present work we are using Pam-250 substitution matrix values. PAM matrice was given by Margaret Dayhoff in 1978. In PAM matrices the PAM-1 matrix gives substitution values which are derived after seeing the mutations of one amino acid into another in every hundred amino acids. These substituted values are considered during the evaluation of the matrix table for the alignment of the sequences.

2. METHODOLOGY
As very short sequences can be aligned by hand but the most highly variable or numerous sequences cannot be solely by human effort. Instead human knowledge is applied in constructing the algorithmically computational approaches to align sequences which fall into two categories global alignments and local alignments. The present methodology for the proposed work involves the use of the cluster analysis techniques to compute the alignment scores between the two sequences. As we know that while aligning the sequences the major goal is to reduce the score for the alignment of the different protein sequence. As less the score will be more efficient alignment will be resulted between the sequences. To generate the sequence alignment we needed to enter the sequences which can be entered through the various formats. The various format available are: Plain Text format, FASTA format, GENBANK and Genetic Computer Group format (GCG). Plain Text format was choose for the problem, as it is a simple format and easy to understand. As it is easy to enter by any of the user (in JAVA applet's text boxes) to understand the protein sequence alignment. A strand of amino acids together make a protein sequence. There are 21 amino acids present.

2.1 Global Alignment
The alignment in which every residue in the strand of sequences is attempted on for alignment is called global alignment. It is a dynamic programming method; technique used for global alignment method is Needleman-Wunsch algorithm. The following algorithm is used, Base conditions:

$F(i, 0) = -i$ and $F(0, j) = -j$ for M(i, j)

The F(i, j) is the score here. The recurrence relation:

F (*i, j*) = *max* {F (*i* – 1, *j*) – 1, F (*i, j* – 1) – 1, F (*i* – 1, *j* – 1) + *score* (*i, j*)}

The time-space complexity of the global alignment is *O* (*mn*).

2.2 Local Alignment

Local alignments are the useful for the most dissimilar sequences that contains most suspected regions of similarity within the larger sequence context. The Smith-Waterman algorithm is used for the local alignment it is dynamic based programming. The algorithm used is, Base conditions:

F(i, 0) = 0 and F(0, j) = 0 for M(i, j)

The recurrence relation:

F(i, j) = max{0, F(i – 1, j) – score(S1(i), –) , F(i, j – 1) – score(–, S2(i)), v(i – 1, j – 1) + score(i, j)}

2.3 Affine Gap Penalties

Affine gaps are induced into which the one value is length independent and other is length dependent. It is calculated through the following algorithm. The gap penalty $\Upsilon(g) = -g * d$. Affine gap score penalizes gap extension less than gap opening:

$$\Upsilon(g) = -d-(g-1) * e, \text{ where } e<d.$$

With general gap penalty $\Upsilon(g)$, base conditions:

M(0,0) = 0, I_x (0,0) = I_y (0,0) = -Infinity;

I_x (i,0) = -d-(i-1) e, M(i,0) = I_y (i,0) = -Infinity, for i= 1,…,n and I_y (i,0) = -d-(j-1) e, M(0,j) = I_y (0,j) = -Infinity, for j= 1,…,n.

The recurrence relation:

$$M(\mathrm{i}, \mathrm{j}) = \max\begin{cases} M(\mathrm{i}-1, \mathrm{j}-1) + \mathrm{s}(\mathrm{x_i}, \mathrm{y_i}) \\ \mathrm{I_x}(\mathrm{i}-1, \mathrm{j}-1) + \mathrm{s}(\mathrm{x_i}, \mathrm{y_i}) \\ \mathrm{I_y}(\mathrm{i}-1, \mathrm{j}-1) + \mathrm{s}(\mathrm{x_i}, \mathrm{y_i}) \end{cases}$$

Here, match or mismatch end gaps in X and Y.

$$I_x(\mathrm{i}, \mathrm{j}) = \max\begin{cases} M(\mathrm{i}-1, j) - \mathrm{d} \\ \mathrm{I_x}(\mathrm{i}-1, j) - e \end{cases}$$

Here, begin gap in X and continue gap in X.

$$I_y(\mathrm{i}, \mathrm{j}) = \max\begin{cases} M(\mathrm{i}, j-1) - \mathrm{d} \\ \mathrm{I_y}(\mathrm{i}, j-1) - e \end{cases}$$

Here, begin gap in Y and continue gap in Y

Complexity is $O(n^3)$ with general gap function, can be reduced to $O(n^2)$ if *k* can be bounded.

2.4 Linear Space Alignment

To calculate the alignment using linear space algorithm we need to obtain the alignment only by replacing the traceback pattern. A new matrix is obtained from the previous matrix obtained. Recall the already matrix and maximum score. We use the last row and column of the calculated matrix for aligning. As this reduces the time complexity to *O*(mn). In this alignment divide and conquer method is applied. We calculate the middle column of the matrix. We traceback from the initial point of matrix to the middle point and then from the middle point to the end point of the matrix. The two solutions are considered to obtain the final.

2.5 Repeated Matches

In this match the alignment of whole sequences from one end to another is done. As if alignment of the sequences is left in between, then its again started from the value higher than the threshold value; till it aligns the whole sequence. Likewise,

HEAGAWGHEE

HEA . AW –HE .

Here the two sequences are aligned. A dot(.) indicates that there is no match between two sequences. A dash(-) indicates that there is a match between two sequences. The algorithm used to calculate it: Each match's score to be greater than *T*. The alignment score is the sum of the match scores substracted by *T*: *score* (*alignment*) = $\sum_i$ (score (*match*$_i$) – *T*). Base conditions: *F* (0, 0) = 0.

$$F(\mathrm{i}, 0) = \max\begin{cases} F(\mathrm{i}-1, 0) \\ \mathrm{F}(\mathrm{i}-1, \mathrm{j}) - \mathrm{T};\ \mathrm{j} = 1, \ldots., \mathrm{m} \end{cases}$$

Recurrence relation:

$$F(\mathrm{i}, \mathrm{j}) = \max\begin{cases} F(\mathrm{i}, 0) \\ \mathrm{F}(\mathrm{i}-1, \mathrm{j}-1) + s(\mathrm{x_i}, \mathrm{y_j}) \\ \mathrm{F}(\mathrm{i}-1, \mathrm{j}) - \mathrm{d} \\ \mathrm{F}(\mathrm{i}, \mathrm{j}-1) - \mathrm{d} \end{cases}$$

Complexity is still *O*(*nm*). Traceback from (n+1, 0) till it reaches zero and if ends in between then its again started from the value higher than threshold value.

2.6 Overlap Matches

Overlap matches are very similar to the local alignment but with the two conditions; first one, the alignment should start from the left side or the top border and secondly, the alignment should start from the right side or the bottom border. The following algorithm is followed: Base conditions: *F*(*i, j*) = 0 if *i* = 0 or *j* = 0. Recurrence rules like those of the global alignment. Recurrence relation:

$$F(\mathrm{i}, 0) = \max\begin{cases} \mathrm{F}(\mathrm{i}-1, 0) \\ \mathrm{F}(\mathrm{i}-1, \mathrm{j}) - T;\ \mathrm{j}=\mathrm{m} \text{ if } \mathrm{i} \le \mathrm{n}, \text{ otherwise } \mathrm{j}=1,....\mathrm{m} \end{cases}$$

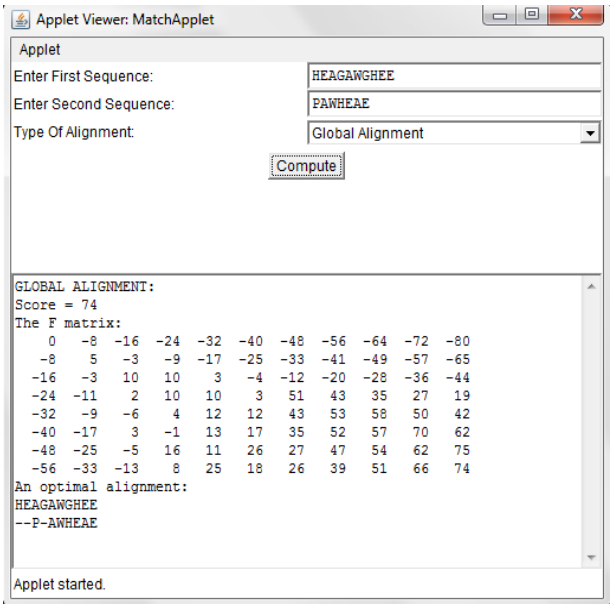
$$F(\mathrm{i}, \mathrm{j}) = \max\begin{cases} \mathrm{F}(\mathrm{i}-1, \mathrm{j}-1) + s(\mathrm{x_i}, \mathrm{y_j}) \\ \mathrm{F}(\mathrm{i}-1, \mathrm{j}) - \mathrm{d} \\ \mathrm{F}(\mathrm{i}, \mathrm{j}-1) - \mathrm{d} \end{cases}$$

3. RESULTS AND DISCUSSION

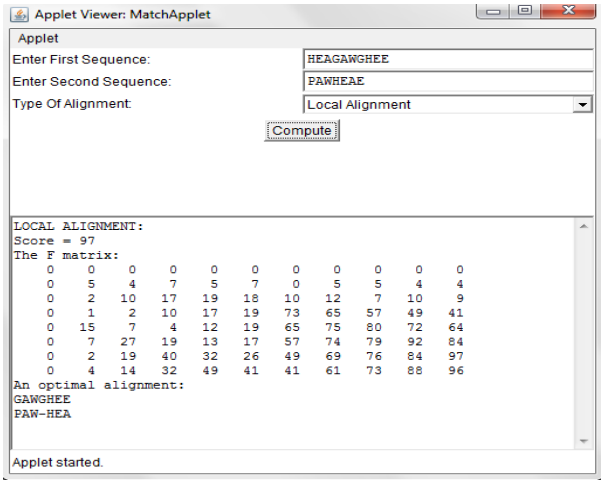
Consider the two sequences. The first sequence is

HEAGAWGHEE and the second sequence is PAWHEAE. Computing alignment using PAM 250 substitution matrix with different algorithms.

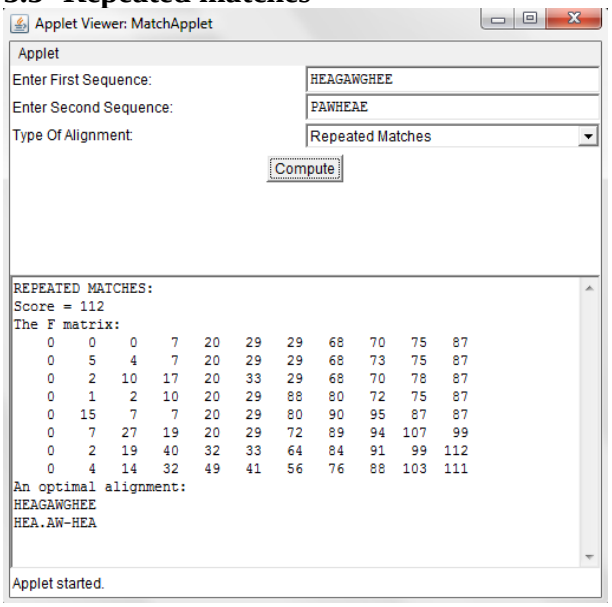
3.1Global Alignment



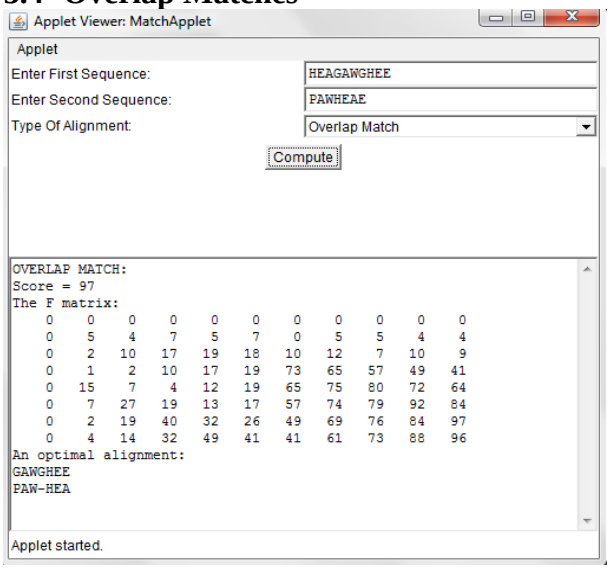
3.2Local Alignment



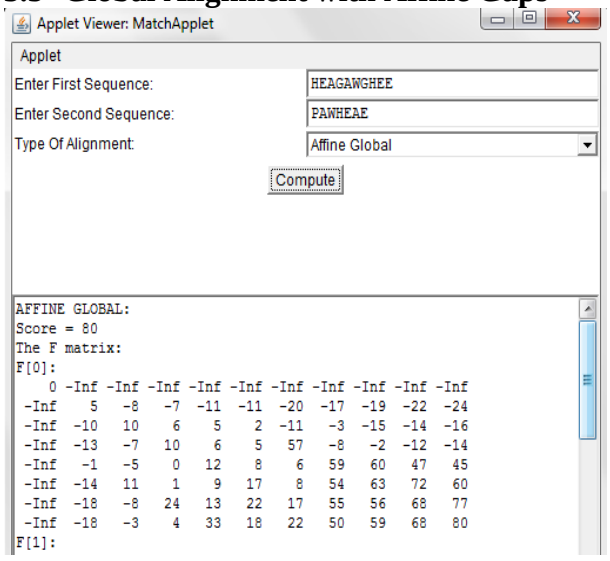
3.3 Repeated matches



3.4 Overlap Matches



3.5 Global Alignment with Affine Gaps

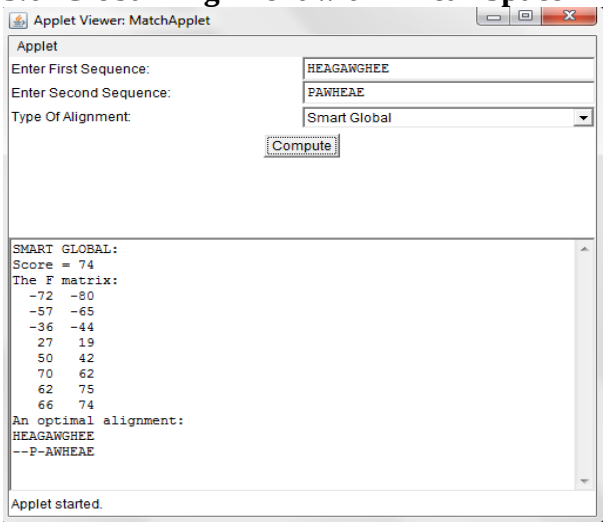


```
F[1]:
    0  -12  -14  -16  -18  -20  -22  -24  -26  -28  -30
 -Inf  -24   -7   -9  -11  -13  -15  -17  -19  -21  -23
 -Inf  -26  -19   -2   -4   -6   -8  -10  -12  -14  -16
 -Inf  -28  -21  -14   -2   -4   -6   45   43   41   39
 -Inf  -30  -13  -15  -12    0   -2   33   47   48   46
 -Inf  -32  -25   -1   -3   -3    5   31   42   51   60
 -Inf  -34  -27  -13   12   10   10   29   43   44   56
 -Inf  -36  -29  -15    0   21   19   27   38   47   56
F[2]:
```

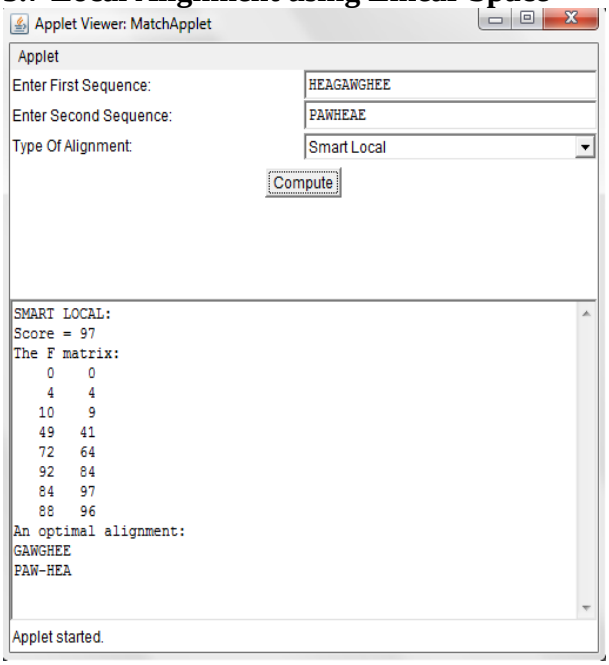
```
F[2]:
    0 -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
  -12  -24  -26  -28  -30  -32  -34  -36  -38  -40  -42
  -14   -7  -19  -19  -23  -23  -27  -29  -31  -33  -35
  -16   -9   -2   -6   -7  -10  -20  -15  -24  -26  -28
  -18  -11   -4   -2   -6   -7   45   33   31   29   27
  -20  -13   -6   -4    0   -4   43   47   48   36   34
  -22  -15   -1   -6   -2    5   41   45   51   60   48
  -24  -17   -3   12    1   10   39   43   49   58   65
An optimal alignment:
HEAGAWGHEE
---PAWHEAE

Applet started.
```

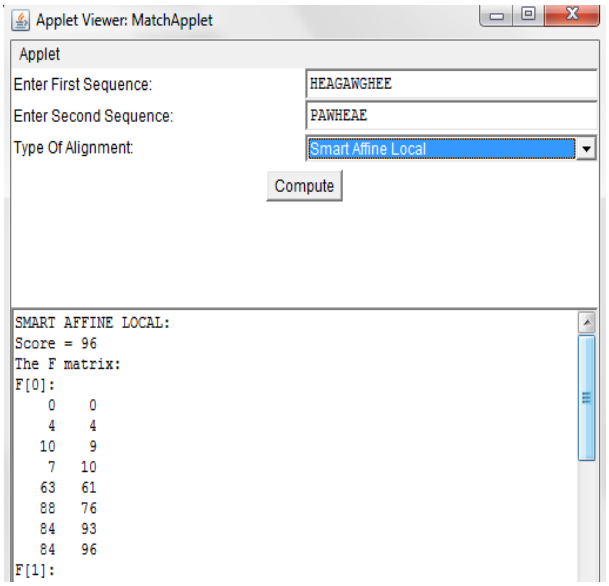
3.6 Global Alignment with Linear Space

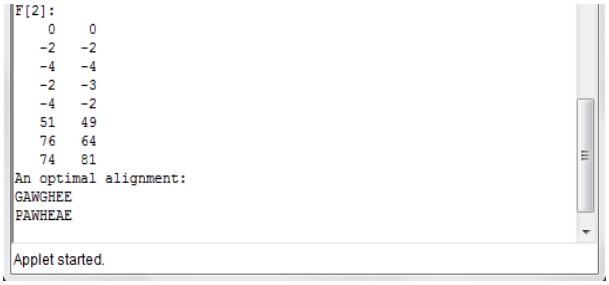


3.7 Local Alignment using Linear Space



3.8 Local Alignment using Linear Space and Affine Gaps





4. CONCLUSIONS

The present work is done in order to construct the different protein sequence alignment by using the substitution values of PAM-250. The protein alignment is useful in predicting the homologous sequences and thus useful for assigning the class for unknown protein. The major reason of using the different alignment scores with the different methods of algorithms is to reduce the score and to get the better alignment between the two sequences. In this way we are able to find more valuable information between the genetically aligned sequences. BLOSUM-50 matrix was used to compare the results with the considered problem using PAM-250 matrix. The substituted model used by different algorithms is done in order to reduce the complexity to compute the alignment between the two sequences. Thus the proposed model is designed in a user friendly option. Thus the user can analyze the different computations steps involved for the final result.

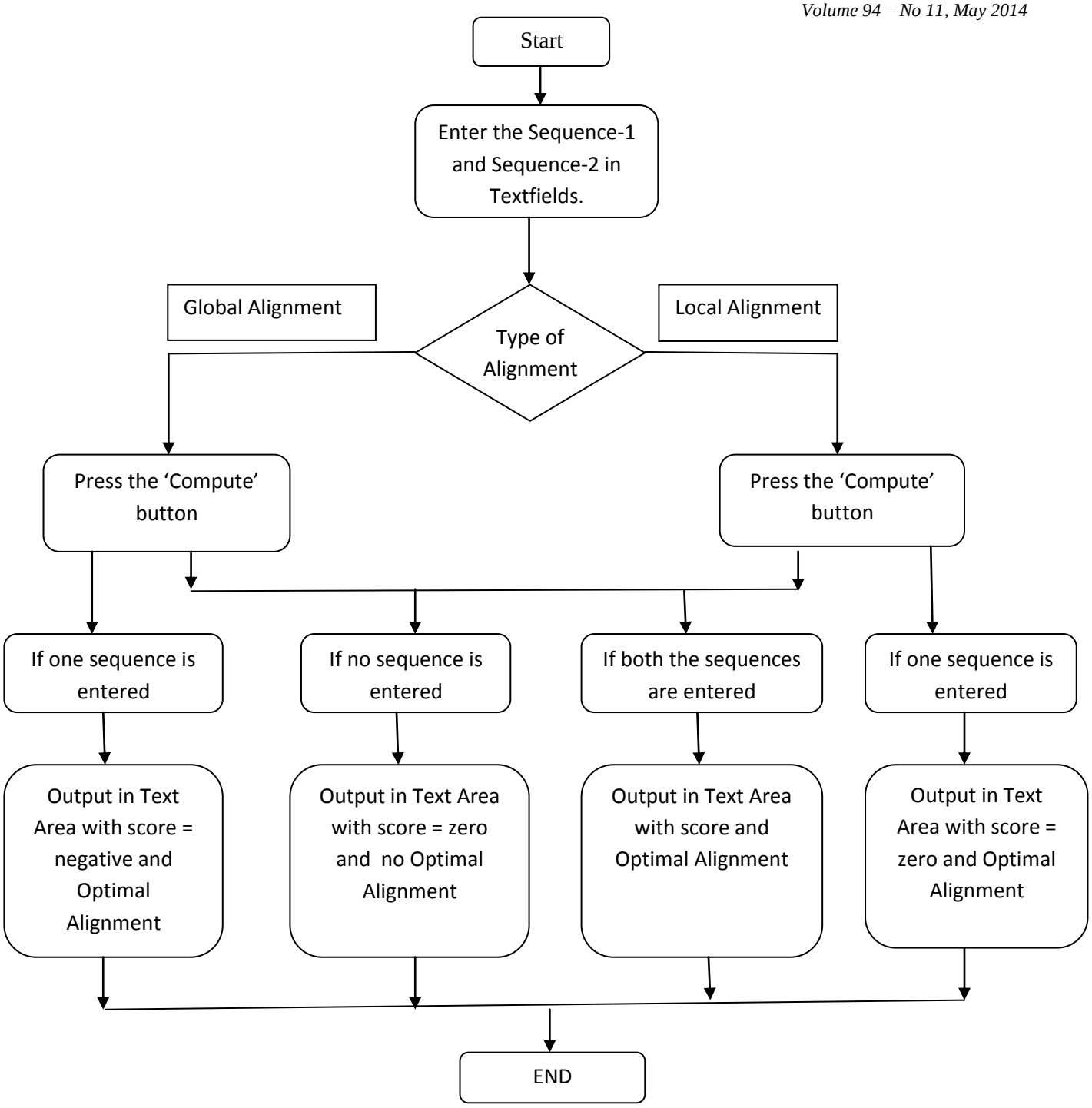


Fig 1:Work Flow of the present work

5. ACKNOWLEDGMENTS

I wish to express my sincere gratitude and indebtedness to my Supervisor, Prof. Rajbir Singh (Assoc. Prof. & Head, Department of Information Technology) for his valuable guidance, attention-grabbing views and obliging nature which led to the successful completion of this study. I lack words to express my cordial thanks to the members of Departmental Research Committee (DRC) for their useful comments and constructive suggestions during all the phases of the present study as well as critically going through the manuscript.

Words fail to express the deep sense of gratitude towards my family members for their moral and financial support and encouragement without which would not have been able to bring out this thesis.

6. REFERENCES

[1] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. .

[2] Pei, J. and Jiang, D. (2005), "*An Interactive Approach to Mining Gene Expression Data",* IEEE Transactions on knowledge and Data Engineering, vol. 17, pp. 1363-1378.

[3] Rastogi, S. C., Mendiratta, N. and Rastogi, P. (2005) "*Bioinformatics Methods and Applications*", third edition, PHI publication, pp.1-350.

[4] Sierk, et al. (2010) *"Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments"*, Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908 USA, pp. 2-15, vol: 146.

[5] Soni, S. Tang, Z. and Yang, J. (2000) "*Performance Study of Microsoft Data Mining Algorithms*", Microsoft White Paper pages 10.

[6] Yehuda, L. (2006), "*Data Mining and Privacy Preserving*", American Association for Artificial Intelligence. Vol. 32, pp 43-54.

[7] Zhang Y, Skolnick J. (2005), "*The protein structure prediction problem could be solved using the current PDB library*". Proc Natl Acad Sci USA 102: 1029–34.

[8] Brick, K. et. al (2008) *"A novel series of compositionally biased substitution matrices for comparing plasmodium proteins"*, Department of Infectious, Parasitic and Immune-Mediated - National Institute of Health, Viale Regina Elena, 299 0016 Roma, Italy.

[9] Sulimova, V. et. al (2008) "Probabilistic evolutionary model for substitution matrices of PAM and BLOSUM families", DIMACS Technical Report 2008-16.

[10] Kantardzic, M. (2000),"Data Mining: Concepts, Models, Methods, and Algorithms", John Wiley & Sons, pp 112-129.

[11] Luscombe, N.M., Greenbaum, D., Gerstein, M. (2001)," *What is Bioinformatics? A proposed definition and Overview of the field*", Luscombe group publications, pp 346-

[12] Merschmann, Luiz and Plastino, Alexandre (2007) "*A Lazy Data Mining Approach for Protein Classification*", Nanobioscience, vol.6, issue 1, March, 2007, pp. 36-42.

[13] Myers, Eugene and Miller, Webb *"Optimal Alignments in Linear Space"*, Department of Computer Science, University of Arizona, Tucson, AZ 85721, NFS Grant DCR-8511455, pp. 1-13

[14] Li, J.J and Huang, De-S. (2005), *"Characterizing Human Gene Splice Sites Using Evolved Regular Expressions"*, Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31 - August 4, 2005, pp. 493-498.

7. AUTHOR'S PROFILE

Rajbir Singh is an Associate Professor & Head, Department of Information Technology of Lala Lajpat Rai Institute of Engineering & Technology Moga India. He received his B.E (Honor) degree in Computer Science and Engineering from MD University, Rothak, Haryana and M-Tech degree in Computer Science and Engineering from Punjab Technical University, Jalandhar Pb. (INDIA). He has authored 03 books on Computer Science. His main field of research interest is Bio-Informatics and Data mining. He works on the Gene Expression, Phylogenetic Trees and Prediction of Protein Sequence & Structure.

Sukhleen Kaur Bhasin is student of M.Tech Department of Computer Science & Engg., Lala Lajpat Rai institute of engg. & Tech, Moga (sukhbhasin@gmail.com), Punjab, INDIA. She received her B.Tech degree in Information and Technology from Punjab Technical University, Jalandhar, Pb. (INDIA). Her research interest includes Bio-Informatics. She worked on the implementation of protein sequence alignment using PAM-250 matrice.

Dheerajpal Kaur is a Faculty with the Department of Electronics & Communication Engineering of Lala Lajpat Rai Institute of Engineering & Technology Moga, India. She received her B.E in Electronics & Communication Engineering and M-Tech degree in Electronics & Communication Engineering from Punjab Technical University, Jalandhar Pb. (INDIA). Her research interests include Neural Networks, Genetics Algorithm and Data Mining. She works on the Antenna Propagation using Neural Networks in MAT Lab