DNA Sequence Alignment in Bioinformatics

[1]Archi Kataria,[2] Dr.Amardeep Singh
[1]*University college of Engineering, Punjabi University, Patiala*
Email: archi.kataria@gmail.com

[2]*University college of Engineering, Punjabi University, Patiala*
Email: amardeep_dhiman@yahoo.com

***Abstract*- In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Computational approaches to sequence alignment generally fall into two categories: *global alignments* and *local alignments.* The two basic alignment algorithms i.e. Smith Waterman for local alignment and Needleman Wunsch for global alignment have been used for sequence alignment. The algorithms have been developed and simulated using MATLAB. The local and global alignment has been performed and the results are shown in the form of Dot plots. For the sequences local and global scores are also calculated and presented**.

Keywords: **Bioinformatics, Biological Sequence Alignment, Smith-Waterman, Needleman-Wunsch, MATLAB, local alignment, global alignment**

I. Introduction

Bioinformatics is an interdisciplinary field that develops and improves upon methods for storing, retrieving, organizing and analyzing biological data. A major activity in bioinformatics is to develop software tools to generate useful biological knowledge. It is a union of biology and informatics as it involves the technology that uses computers for storage, retrieval, manipulation and distribution of information related to biological macromolecules such as DNA, RNA and proteins[1].
Mapping and analyzing DNA and protein sequences, aligning different DNA and protein sequences to compare them and creating and viewing 3-D models of protein structures are common activities in Bioinformatics. Major research efforts in the field includes sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, genome-wise association studies and modeling of association[7].
Biological sequence alignment aims to find out whether two or more biological sequences i.e. DNA, RNA, or Protein sequences are related or not and is very important and widely used operation in the field of bioinformatics and computational biology. This has many important real world applications as if some information about one of the sequences is already known (e.g., the sequence represents a cancerous gene) then this information could be transferred to the other unknown sequences, which could be vital in early disease diagnosis and drug engineering [2].

A. Sequence Alignment

Sequence alignment is the process by which sequences are compared by searching for common character patterns and establishing residue-residue correspondence among related sequences [3]. Homologous sequences may have different length, though, which is generally explained through insertions or deletions in sequences. In genome projects, newly determined sequences are first compared with those placed in genomic databases in order to discover similarities. This is done because relevant sequence similarity is evidence of common evolutionary origin and homology relationship. A sequence alignment has a similarity score associated to it that is obtained by placing one sequence above the other. The rapid evolution of sequencing techniques combined with the intense growth in the number of large-scale genome projects is producing a huge amount of biological sequence data. Nevertheless, determining the genome sequence is only the first step toward deciphering the genetic message encoded in those sequences [4].

II. Alignment Methods and Algorithms

Computational approaches to sequence alignment generally fall into two categories: *global alignment* and *local alignment.* The length of a sequence is variable and sometimes we require the alignment of lengthy and highly variable or extremely numerous sequences. Constructing algorithms to produce high-quality sequence alignments using four letters becomes a real challenge By global alignment, we consider aligning the entire scope of all

© 2013 IJAIR. ALL RIGHTS RESERVED

query sequences against a reference sequence. On the other hand, the method of local alignment identifies isolated regions of high similarity within the entire sequence, which makes the technique a better choice in some situations but a more complex one in general[7].

A. Global Alignment

An alignment that assumes that the two proteins are basically similar over the entire length of one another. The alignment attempts to match them to each other from end to end, even though parts of the alignment are not very convincing.[9]Needleman-Wunsch algorithm is used for Global alignment.

a. Needleman-Wunsch (Global Alignment)

The Needleman-Wunsch algorithm is a dynamic programming algorithm for optimal sequence alignment (Needleman and Wunsch, 1970). Basically, the concept behind the Needleman-Wunsch algorithm stems from the observation that any partial sub-path that tends at a point along the true optimal path must itself be the optimal path leading up to that point. Therefore the optimal path can be determined by incremental extension of the optimal sub-paths. In a Needleman- Wunsch alignment, the optimal path must stretch from beginning to end in both sequences (hence the term 'global alignment').

In order to perform a Needleman-Wunsch alignment, a matrix is created which allows us to compare the two sequences. The score M(i,j) for every cell depends on the three cells corresponding to either or both sequence having 1 less letter (i.e. cells M(i-1.j), M(i,j-1) and M(i-1,j-1). It is calculated as follows:

$$M(i,j) = MAX(Mi-1,j-1 + S(Ai, Bj)$$
$$Mi-1, j + gap$$
$$Mi,j-1 + gap)$$

where gap is the gap penalty and the function S returns the score/penalty for matching the two corresponding letters. Once we have computed this score for every cell, we must do a "traceback", that is to determine the actual set of operations that lead to the score.

B. Local Alignment

An alignment that searches for segments of the two sequences that match well. There is no attempt to force entire sequences into an alignment, just those parts that appear to have good similarity.[9] Smith-Waterman algorithm is used for local alignment.

b. Smith-Waterman (Local Alignment)

Smith-Waterman algorithm calculates the local alignment of two given sequences. It is used to identify similar DNA, RNA and protein segments. The alignments of any possible length starting and ending at any position in the two sequences are compared to obtain the optimal local alignment. In this algorithm, the alignment path does not need to reach the edges of the search graph,but may begin and end internally. In order to accomplish this, 0 was added as a term in the score calculation described by Needleman and Wunsch. It is calculated as follows:

$$M(I,j) = MAX(Mi-1,j-1 + S(Ai, Bj)$$
$$Mi-1,j + gap$$
$$Mi,j-1 + gap$$
$$0)$$

The implication of this is that there are no values below zero in a local alignment scoring matrix[10].

III. DNA Sequence Alignment

For applying a global and local alignment and to get a score for both of them, the user can enter the sequence in two ways. The first way is by the accession numbers of the sequence to retrieve the sequences in its ORF (Open Reading Frames). The second way is to retrieve the sequences from the web (public database) and bringing the sequence information into the MATLAB environment [5]. After that we can get global alignment (NW) and local alignment (SW) with a score that determines the degree of similarity. Dot plots are one of the easiest ways to find similarity between the two sequences. Many dots in the dot plot line up to form diagonal lines indicating good alignment between the two sequences.

IV. Experimental Results and Discussion

A. To retrieve sequences from a database:-

Different sequences that have to be analyzed, aligned and read are retrieved from public database into MATLAB environment.

© 2013 IJAIR. ALL RIGHTS RESERVED

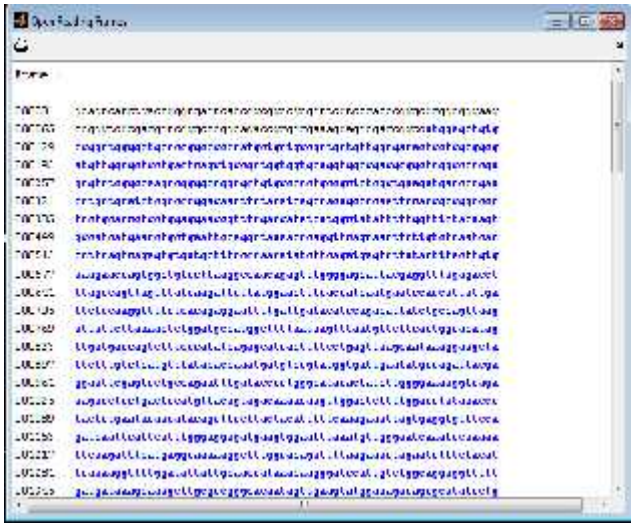

Fig 1 Open Reading Frame of Human DNA sequence

B. Sequence comparison by using dot plot:-

The most basic sequence alignment method is the dot matrix method also known as dot plot method. It is the graphical way of comparing two dimensional matrix [6].In dot matrix plot, the two sequences to be compared are written along the top row and leftmost column of a two-dimensional matrix and a dot is placed at any point where the characters in the appropriate columns match[8]. When the two sequences have substantial regions of similarity, many dots line up to form diagonal lines, which reveals the sequence alignment.

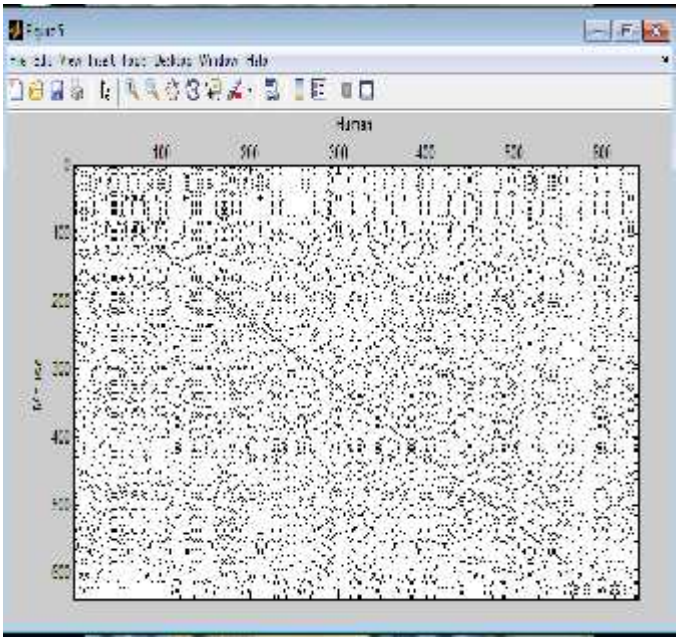

Fig 2 Dot plot of Human and Mouse

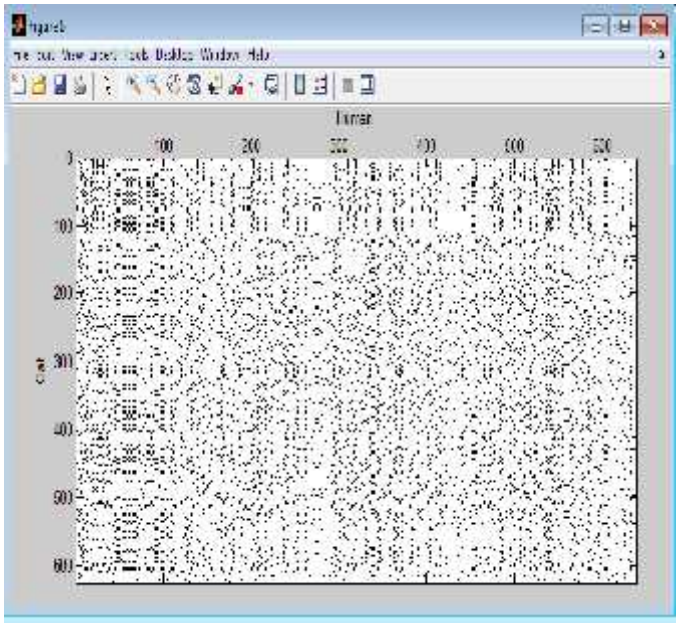

Fig 3 Dot plot of Human and Cat

Dot plots of human and mouse DNA sequences and human and cat DNA sequences have been shown in Fig. 2 and Fig. 3 respectively. The dot plots above shown shows that human and mouse DNA sequences show better alignment as compared to human and cat DNA sequences.

C. Global alignment of sequences:-

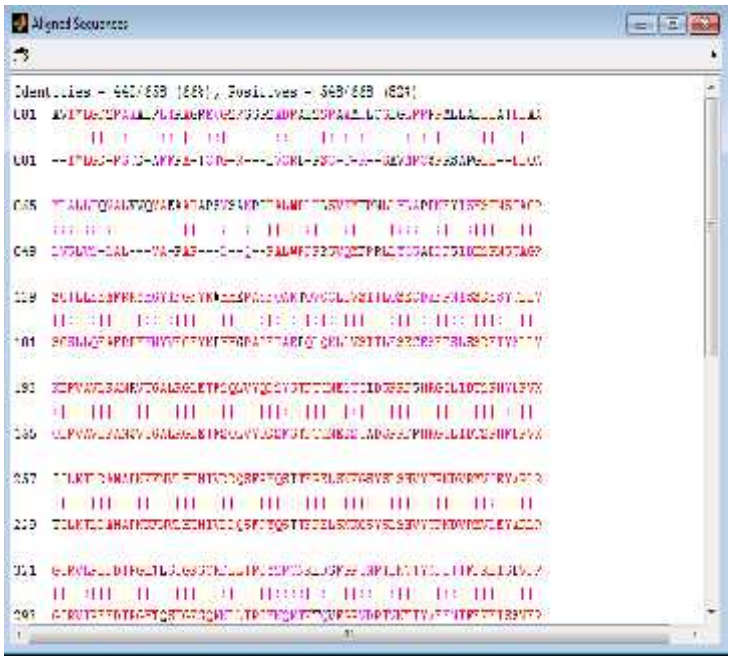

Fig 4 Global Alignment of Human and Mouse

© 2013 IJAIR. ALL RIGHTS RESERVED

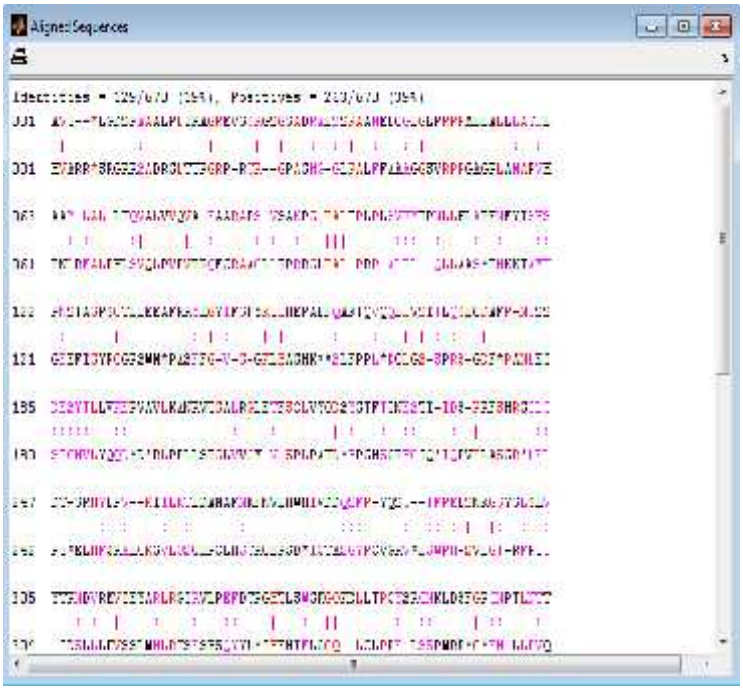

Fig 5 Global Alignment of Human and Cat

Global alignment of human and Mouse sequences is shown in Fig. 4 and of human and Cat sequences is shown in Fig.5 respectively.

D. Local alignment of sequences:-

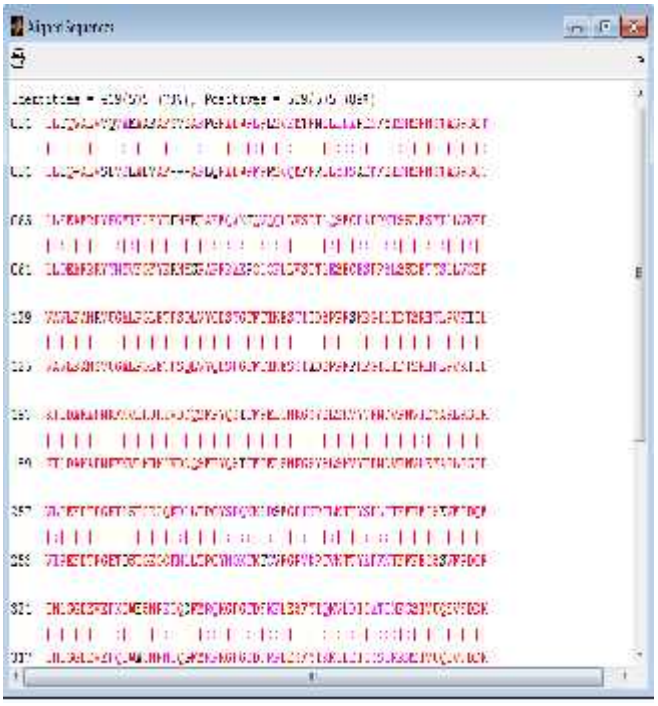

Fig 6 Local Alignment of Human and Mouse

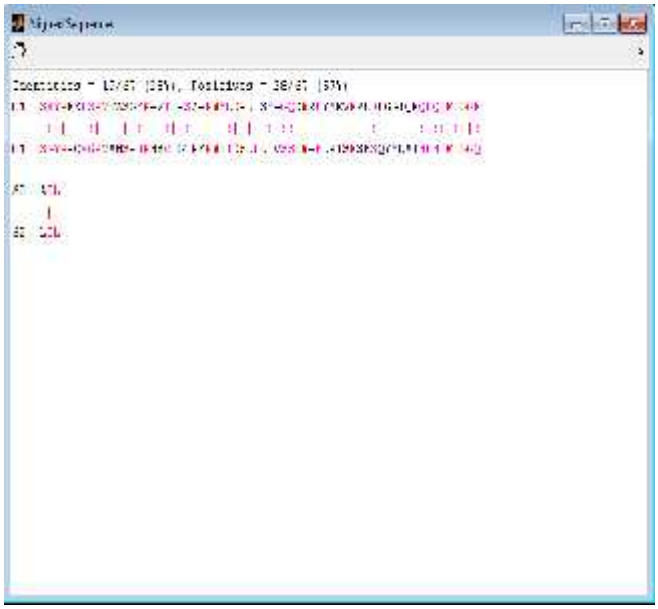

Fig 7 Local Alignment of Human and Cat

Local alignment of human and Mouse sequences is shown in Fig.6 and of human and Cat sequences is shown in Fig.7 respectively.

V. Conclusion

Sequence alignments algorithm have been developed and simulated using MATLAB. Sequence alignment results have been presented in the form of dot plots, local alignment score and global alignment score. The alignment score for human and Mouse sequences for global alignment is 875 and for local alignment it is 947.66. The alignment score for human and Cat sequences is -99.6667 for global alignment and is 21 for local alignment.

References

[1] Hassan Mathkour, Muneer Ahmad "A Comprehensive Survey on Genome Sequence Analysis"*IEEE International Conference on Bioinformatics and Biomedical Technology*, pp. 14-18, 2010

[2] Changjin Hong, Ahmed H. Tewfic "Heuristic Reusable Dynamic Programming: Efficient Updates of Local Sequence Alignment*" IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 6, No. 4, pp. 570-562, 2009

[3] Khaled Benkrid, Ying Liu, AbdSamad Benkrid "A Highly Parameterized and Efficient FPGA-Based Skeleton for Pairwise

© 2013 IJAIR. ALL RIGHTS RESERVED

Biological Sequence Alignment*" IEEE Transactions on Very Large Scale Integration Systems*, Vol. 17, No. 4, pp. 561-570, 2009

[4] Azzedine Boukerche, Jan M. Correa, Alba Cristina M.A de Melo, Ricardo P. Jacobi "A Hardware Accelerator for the Fast Retrieval of DIALIGN Biological Sequence Alignments in Linear Space*" IEEE Transactions on Computers*, Vol. 59, No. 6, pp. 808-821, 2010

[5] Nahar, N.L. Hamel, M.S. Popstova and J.P. Gogarten "GPX: A Tool for the Exploration and visualization of Genome Evolution" *IEEE International Conference on Bioinformatics and Bioengineering*, pp. 1338 – 1342, 2007

[6] Mai S.Mabrouk, Marva Hamdy, MarvaMamdouh, Marva Aboelfotoh, Yesser M.Kadah "BIOINFTool: Bioinformatics and sequence data analysis in molecular biology using Matlab" *Cairo International Biomedical Engineering Conference*, pp. 1-9, 2006

[7] Sonali Vijan and Rajesh Mehra "Biological Sequence Alignment for Bioinformatics Applications Using MATLAB" *Int. J Comp Sci. Emerging Tech*, Vol-2, No 5, October, 2011

Web references

[8] http://en.wikipedia.org/wiki/Sequence Alignment

[9] http://www.ebi.ac.uk/2can/tutorials/protein/align.html

[10] http://www.cs.utoronto.ca/~brudno/bcb410/lec2notes.pdf

© 2013 IJAIR. ALL RIGHTS RESERVED