ISSN:2231-0711
www.ijcset.net

DNA Sequence Alignment based on Bioinformatics

Shivani Sharma, Amardeep singh

Computer Engineering,Punjabi University,Patiala,India
Email: Shivanisharma89@hotmail.com

Abstract: **DNA Sequence alignmentis the most fundamental and essential task of computational biology and forms the base for other tasks of bioinformatics. The two basic alignment algorithms i.e. Smith Waterman for local alignment and Needleman Wunsch for global alignment have been used in this paper. The algorithms have been developed and simulated using MATLAB for genome analysis and sequence alignment. The local and global alignment has been presented and the results are shown in the form of Dot plots and local and global scores for the sequences.**

Keywords:Bioinformatics, DNA Sequence Alignment, Smith-Waterman, Needleman-Wunsch, local alignment, global alignment

1. INTRODUCTION

All living organism cells are composed of genetic codes thatare passed from one generation to other. This is the reason forsome living organisms being biologically similar and somebeing distinct. The genetic code can be represented as asequence of alphabets, such as four base pairs of DNA andRNA, or twenty amino acids of protein[1].These sequences are called biological sequences and over time a lot of changescalled mutations occur in these sequences.

The field of bioinformatics aims to align a large number of biological sequences with the purpose of deriving their evolutionary relationships through comparative sequence analysis. The bioinformatics applies computations to the biological sequences in order to analyze and manipulate them.Common activities in Bioinformatics include 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.

Sequence alignment is the most basic and essential module of computational bioinformatics and has varied applications in sequence assembly, sequence annotation, structural and functional prediction, evolutionary or phylogeny relationship analysis. It aims to find out whether two or more biological sequences are related or not.In biomolecular sequences (DNA, RNA, or amino acid sequences) high sequence similarity usually implies significant functional or structural similarity.

2. SEQUENCE ALIGNMENT

Any biological sequence is a sequence of characters drawn from an alphabet. For DNA sequence, character set is {A, C, G, T}, for RNA sequence, the set is {A, C, G, U}, and for protein sequence, character set is {A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V}. A sequence alignment is the process of identifying one-to-one correspondence among subunits of sequences in order to measure the similarities among them[3]**.**

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. 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. Sequence comparison is a very basic but important step in genome projects. As a result of this, one or more sequence alignments can be produced. A sequence alignment has a similarity score associated to it that is obtained by placing one sequence above the other making clear the correspondence between the characters. Two approaches to sequence alignment are:-

2.1 Global Alignment

Global alignment get the maximum match between the sequences as it assume that the two sequences are similar. This alignment attempts to match the two sequences from the end to the end even though if they are different in some parts.Sequences that are quite similar and approximately the same length are suitable candidates for global alignment[2].

```
NLGPSTKDFGKISESREFDNQ
  |          | | | |          |
QLNQLERSFGKINMRLEDALV
```

Fig 1: Global Alignment of two sequences

2.2 Local Alignment

Local alignment searches for the part of the two sequences that match well.In this, stretches of sequencewith the highest density of matches are aligned, thus generating one or more islands of matches or subalignments in the aligned sequences. Local alignments are more suitable for aligning sequences that are similar along some of their lengths but dissimilar in others, sequences that differ in length orsequences that share a conserved region or domain[1].

NLGPSTKDDFGKILGPSTKDDQ
| | ||
QNQLERSSNFGKINQLERSSNN

Fig 2: Local Alignment of two sequences

3. Alignment Algorithms

DNA sequences are strings of letters from a four-letter alphabet called nucleotides (A, C, G, T). The length of a sequence is variable and sometimes we require the alignment of lengthy and highly variable or extremely numerous sequences. Hence, constructing algorithms to produce high-quality sequence alignments using four letters becomes a real challenge. In general two types of sequence alignment are classified as local and global alignment.Local alignments identify regions of similarity within long sequences that are often widely divergentoverall. Global alignment forces the alignment tospan the entire length of all query sequences.

3.1 Needleman-Wunsch Algorithm

The Needleman-Wunschalgorithm[4] performs a global alignment on two query sequences and is used widely in bioinformatics to align protein or nucleotide sequences. It uses a dynamic programming method to ensure the alignment is optimum by exploring all possible alignments and choosing the best.

In Needleman-Wunsch Algorithm, first take the two sequences and create a 2-dimensional array with the length of multiply of the two sequence's length, each cell can be evaluated from the maximum of the three cells around it and at the same time keep a pointer of the maximum value to make the trace-back to get optimal solution[6][7]. The steps of the Needleman-Wunsch algorithm is as follows:-

Input: two sequences x and y
Output: optimal alignment and score
Initialization:
Set G(0,0)=0
Set G(i, 0) := −id for all i = 0, 1, 2, . . . , m
Set G(0, j) := −jd for all j = 0, 1, 2, . . . ,n
Main Iteration:
Filling in partial alignment
For each i = 1, 2, . . . , m do:
For eachj = 1, 2, . . . ,n do:
Set G(i, j)= max{G(i − 1, j − 1) + s(xi, yj),case 1
G(i − 1, j) + d, case2
G(i, j − 1) + d, case3}

Ptr(i, j) = { DIAG , if case 1
LEFT , if case 2
UP , if case 3 }

Termination:
G(M, N) is the optimal score, and from Ptr(M, N),we can trace back optimal alignment.

3.2 Smith Waterman Algorithm

The Smith–Waterman algorithm is a well-known algorithm for performing local sequence alignment that is for determining similar regions between two nucleotide or protein sequences[5][9]. Instead of looking at the total sequence, the Smith–Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure.The main difference is that G[i, j] add to the maximum function that declared in Needleman and Wunsch the possibility of Zero value. The formula for computing G[i, j] becomes:

G(i, j) = max { 0;
G(i −1, j −1) + s(xi,yj);
G(i −1, j) + d;
G(i, j −1) + d}

4. Proposed DNA Sequence Alignment

For applying a global and local alignment and getting 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 Open Reading Frames. The second way is to retrieve the sequences from the web and bringing the sequence information into the MATLAB environment.After that we can get global alignment and local alignment with a score that determines the degree of similarity. Dotplots are one of the easiest ways to look for similaritybetween sequences. The diagonal line indicates that there may be a good alignment betweenthe two sequences.

5. Results and Discussion

5.1 *Retrieve sequences from a database:-*

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

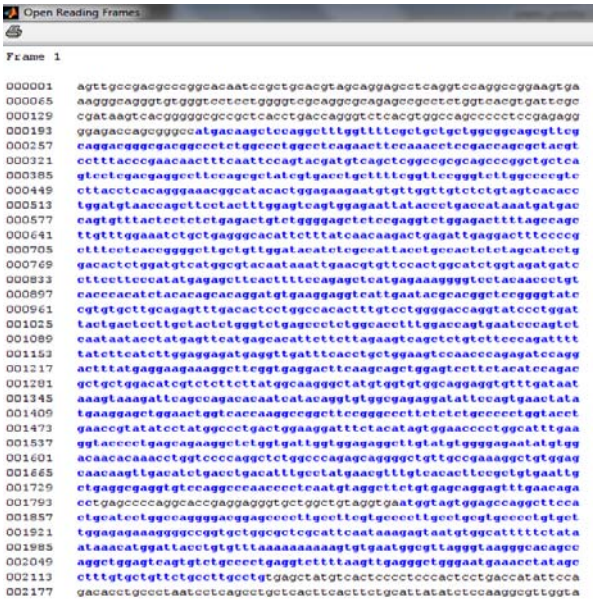

Fig 3 Open Reading Frame of Human DNA sequence

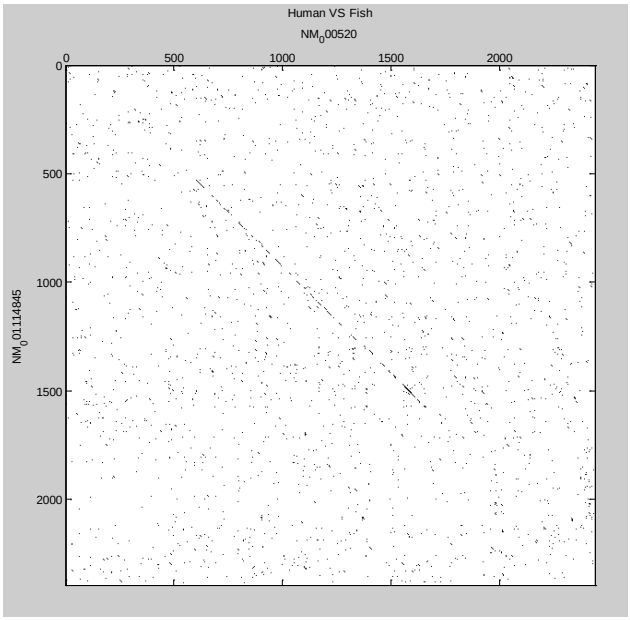

Fig 5 Dot plot of Human and Zebrafish

5.2 Sequence comparison by using dot matrix method:-

The most basic sequence alignment method is the dot matrix method also known as dot plot method.In dotmatrix 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]. The dot plots of very closely related sequences will appear as a single line along the matrix's main diagonal.MATLAB function has been used for this comparison.

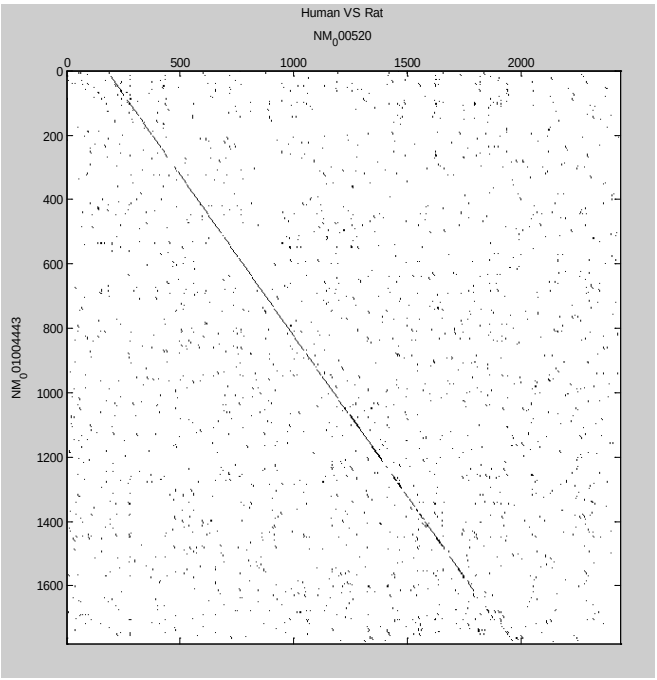

Fig 4 Dot plot of Human and Norway Rat

Dot plots of human and norway rat DNA sequences and human and zebrafish DNA sequences have been shown in fig. 4 and fig. 5 respectively. The dot plots above shown shows that human and Norway rat DNA sequences show better alignment as compared to human and zebrafish DNA sequences.

5.3 Global alignment of sequences by using Needleman Wunsch Algorithm

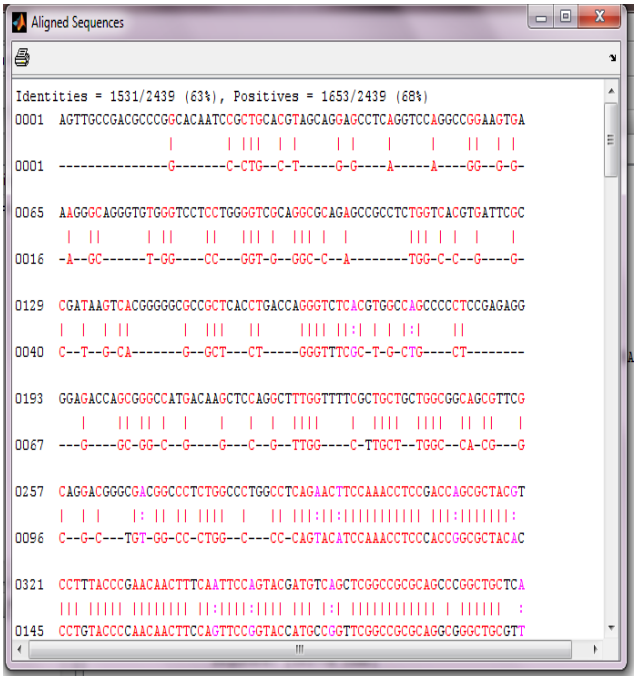

Fig 6 Global Alignment (NW) of Human and Norway Rat

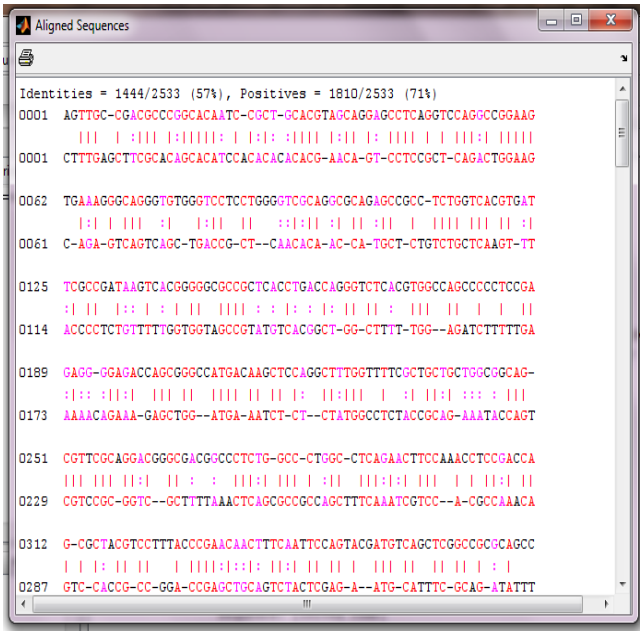

Fig 7 Global Alignment (NW) of Human and Zebrafish

Global alignment of human and Norway rat DNA sequences is shown in fig. 6 and of human and zebrafish DNA sequences is shown in fig.7 respectively.The alignment score for human and norway rat DNA sequences is 2287 and for human and zebrafish DNA sequences is 2893 for global alignment.

5.4 Local alignment of sequences by using Smith Waterman Algorithm

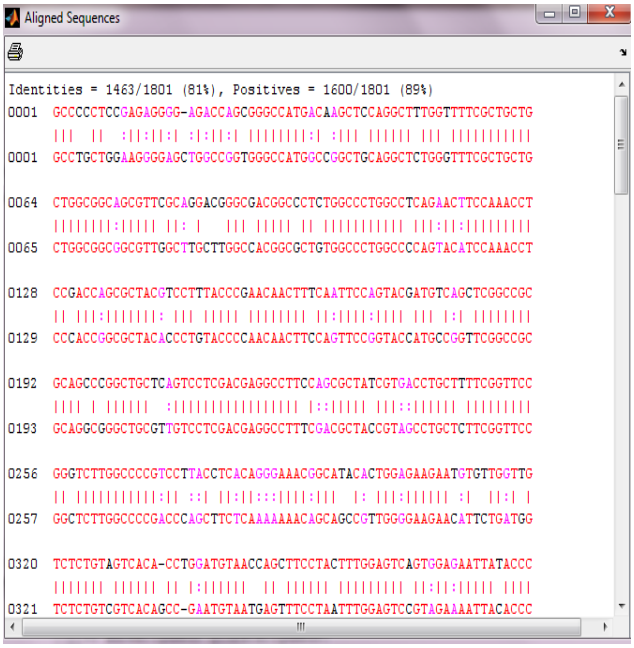

Fig 8 Local Alignment (SW) of Human and Norway rat

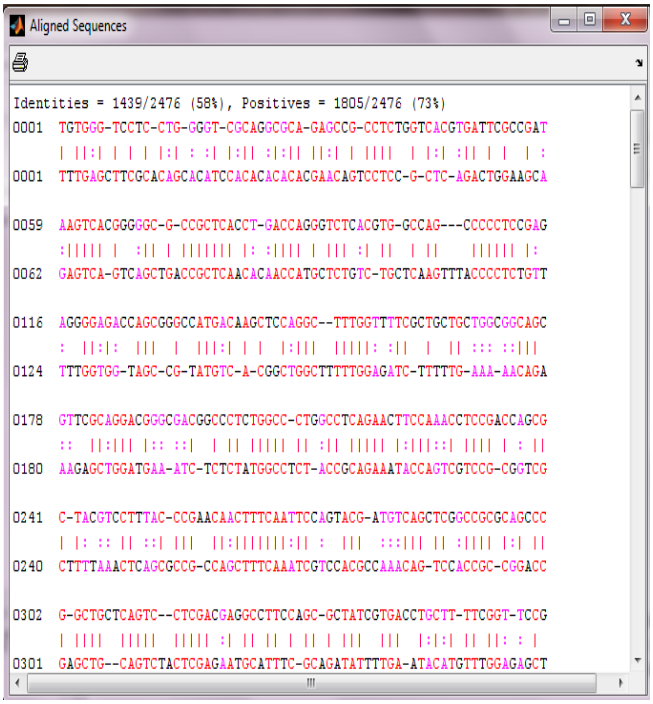

Fig 9 Local Alignment (SW) of Human and Zebrafish

Local alignment of human and Norway rat DNA sequences is shown in fig.8 and of human and zebrafish DNA sequences is shown in fig.9.The alignment score for human and norway rat DNA sequences is 3750 and for human and zebrafish DNA sequences is 2.9567e+003 for local alignment.

6. CONCLUSION

In this paper DNA sequence alignments algorithm have been developed and simulated using MATLAB. MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy to use environment where problems and solutions areexpressed in familiar mathematical notation.Sequence alignment results have been presented in the form of dot plots, local alignment score by using Smith Waterman algorithm and global alignment score by using Needleman Wunsch algorithm. The alignment score for human and Norway rat DNA sequences for global alignment is 2287 and for local alignment it is 3750. The alignment score for human and zebrafish DNA sequences is 2893 for global alignment and is 2.9567e+003 for local alignment. The proposed work is a useful tool that can aid in the exploration, interpretation and visualization of data in the field of molecular biology.

REFERENCES

[1] Mai S.Mabrouk, MarvaHamdy, MarvaMamdouh, MarvaAboelfotoh, YesserM.Kadah, *"BIOINFTool: Bioinformatics and sequence data analysis in molecular biology using Matlab"*, Cairo International Biomedical Engineering Conference, pp. 1-9, 2006

[2] TahirNaveed, ImitazSaeedSiddiqui, Shaftab Ahmed, "Parallel Needleman-Wunsch Algorithm for Grid", Proceedings of the PAK-US International Symposium on High Capacity Optical Networks and Enabling Technologies (HONET 2005), Islamabad, Pakistan, Dec 19 - 21, 2005

[3] Nivit Gill, Shailendra Singh, ***"****Multiple Sequence Alignment using Boolean Algebra and Fuzzy Logic: A Comparative Study"* Int. J. Comp. Tech. Appl., Vol 2 (5), 1145-1152, 2011

[4] Needleman S, Wunsch.,"A general method applicable to the search for similarities in the amino acid sequences of two proteins", Journal of Molecular Biology 48, 443-453, 1970.

[5] SérgioAnibal de CarvalhoJunior , "Sequence Alignment Algorithms", thesis, 2002/2003**.**

[6]SaraA.Shehab, ArabiKeshk, HanyMahgoub, "Fast Dynamic Algorithm for Sequence Alignment based on Bioinformatics",International Journal of Computer Applications (0975 – 8887) Volume 37– No.7, 2012

[7] Wikipedia Website, "Needleman-WunschAlgorithm",http://en.wikipedia.org/wiki/Needleman-Wunsch algorithm.

[8] Wikipedia Website,"Sequence Alignment", http://en.wikipedia.org/wiki/Sequence Alignment

[9] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequence," *J. Molecular Biology*, vol. 147, pp. 195-197, 1981.

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

[11] Source of DNA Sequences (online), National Center for Biotechnology Information.