EUROPEAN JOURNAL OF BIOMEDICAL AND PHARMACEUTICAL SCIENCES
http://www.ejbps.com

ISSN 2349-8870
Volume: 6
Issue: 5
134-142
Year: 2019

SEQUENCE ALIGNMENT

[1]Shireen Begum and [2]*Saba Yousuf

[1]Assistant Professor, Deccan School of Pharmacy
[2]Student, Deccan School of Pharmacy (Affiliated to O.U)*
Department of Pharmaceutics, Deccan School of Pharmacy, Darussalam, Aghapura, Hyderabad-500001, Telangana, India.

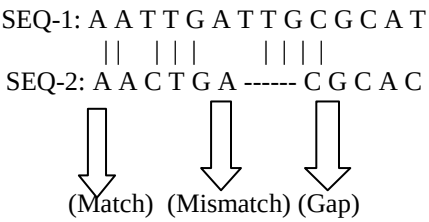
***Corresponding Author: Saba Yousuf**
Student, Deccan School of Pharmacy (Affiliated to O.U), Department of Pharmaceutics, Deccan School of Pharmacy, Darussalam, Aghapura, Hyderabad-500001, Telangana, India. **Email**

Article Received on 28/02/2019 **Article Revised on 18/03/2019** **Article Accepted on 08/04/2019**

ABSTRACT
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. The number of identical and similar amino acid residues may be compared to the total number of amino acids in the protein. This gives the percentage of identical and similar residues-percentage of sequence identity and sequence similarity. Similar residues are those that have similar chemical characteristics. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns. The important models for sequence analysis are GLOBAL ALIGNMENT, LOCAL ALIGNMENT and GAP PENALTY. The Smith waterman algorithm is a local alignment algorithm which is able to identify mutation in DNA sequences. The various methods for sequence alignment include BRUTE FORCE, DOT MATRIX, DYNAMIC PROGRAMMING, HEURISTICS METHODS. Scoring matrices are used for computing alignment scores, based on observed substitution rates derived from the substitution frequencies. The most popular two scoring matrices are PAM and BLOSUM matrices.

KEYWORDS: Global alignment, Local alignment, Heuristics methods, Scoring matrices, PAM and BLOSUM matrices.

DEFINITION
Sequence alignment is the procedure of comparing two (pair wise) or more (multiple sequence alignment) sequences of DNA/RNA or proteins to identify the similarity which may be the consequence of structural, functional or evolutionary relationship by searching for a series of individual characters or characters patterns that are in the same order in the sequences.[1]



An alignment of two short DNA sequences

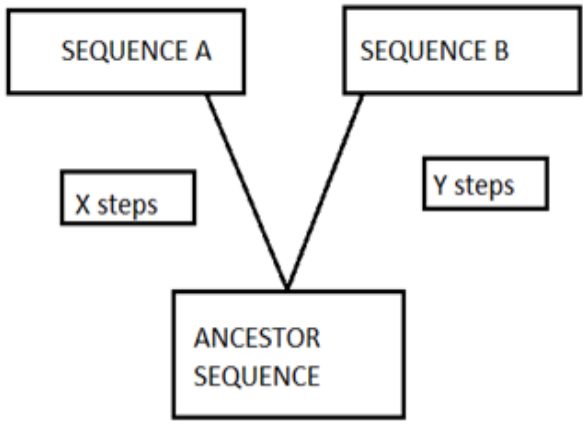
REPRESENTATION
Two sequences are aligned by writing them across a page in two rows. Identical or similar characters are placed in the same column, and non- identical characters can either be placed in the same column as a mis- match or opposite a gap in the other sequence. In an optimal alignment, non-identical characters and gaps are placed to bring as many identical or similar characters as possible into vertical register.

In the above figure TTG (letters 7-9) are not aligned with any letters from sequence-2. When this happens, we say that a gap has been introduced. The point of introducing a gap here is that it enables a letter alignment of two sequences, in which more of the aligned pairs are identical letters. In this case, it enables the shared CGCA directly after the gap to align between two sequences Alignment can be interpreted in evolutionary term. When identical letters are aligned, the simplest interpretation is that these letters were part of the ancestral sequence and have remained unchanged. When non identical letters are aligned, the simplest interpreted in terms of the insertion or deletion of letters in one of the sequences. Gap is sometimes referred as indel Without the ancestral

sequence, it is not possible to know in which sequence the mutation or gaps actually occurred.[1]

SIGNIFICANCE

- Sequence alignment is useful for discovering functional, structural and evolutionary information in biological sequence.
- It is important to obtain the best possible or so called optimal alignment to discover the information sequences which are very alike or similar in the parlance of the sequence analysis, probably have the same function be it a regulatory role in the case of similar DNA molecules, or a similar biochemical function and three dimensional structure in the case of proteins.
- Additionally if two sequences from different organisms are similar, there may be a common ancestor sequence, and the sequences are then defined as homologous. The alignment indicates the changes that could have been occurred between the two homologous sequences and a common ancestor during evolution.[1]



Ancestor sequence

MODELS FOR SEQUENCE ANALYSIS[1],[2]

The important models for sequential analysis are described here:

1. Global alignment
2. Local alignment
3. Gap penalty

1. GLOBAL ALIGNMENT

Global alignment is done across the entire sequence length to include as many as matches as possible up to and including the sequence end. In global alignment an attempt is made to align the entire sequence using as many characters as possible up to both ends of each sequence (Needleman wunsch).

L G P S K T G K G S – S R I
| ‖ | | | |
L N – I K A G K G A I M R L

Global Alignment

2. LOCAL ALIGNMENT: The task is to find and extract pair of region, one from each of the two given strings that exhibit high similarity. This is called the local alignment or local similarity problem (smith waterman).

----------T G K G----------
| | |
----------A G K G----------

Local Alignments

Global alignment v/s Local alignment

GLOBAL ALIGNMENT	LOCAL ALIGNMENT
• Global Alignment is stretched over the entire sequence length to include many matching of amino acid or nucleotide as possible up to and including the sequence ends. • Sequences that are quite similar and approximately the same length are suitable candidates for global alignment	• In local Alignment, the alignment stops at the ends of the regions of identity or strong similarity, higher priority to find these regions. • This type of alignment favors finding conserved nucleotide patterns, DNA sequences or amino acids patterns in protein sequence.

3. GAP PENALTY

A gap is any maximal, consecutive run of spaces in a single string of a given alignment. Gaps help create alignments that better conform to underlying biological models and more closely fit patterns that one expects to find in meaningful alignment. The idea is to take in account the number of continuous gaps and not only the number of spaces when calculating an alignment.

For example consider the alignment:
S = a t t c - - g a – t g g a c c
T = a - - c g t g a t t - - - c c

METHODS OF SEQUENCE ANALYSIS[1],[3]

These are the various methods of sequence alignment:

1. BRUTE FORCE

Brute force method is based on exhaustive enumeration. It produces alignments without gaps and has a N2 complexity where N is the length of the sequences. This is a trivial method with hardly an Y practical utility.

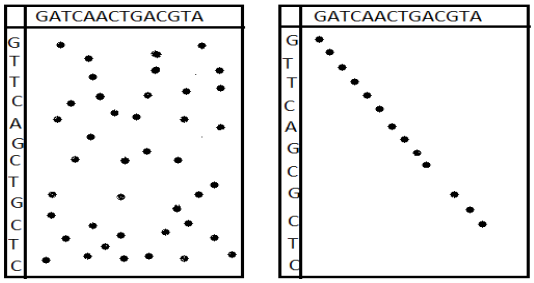
2. DOT MATRIX

A dot matrix analysis is a method for comparing two sequences and tries to locate possible alignment of characters between the two sequences. Similarity between two sequences can be detected as a diagonal on an identity matrix. Graphic similarity comparisons use the power of the computer to present the relationships between sequences in such a graphic form that enables the human researcher to discern patterns in the data. Dot matrix method is useful for simple alignments. It utilizes graphical methods, which are easy to understand and apply. However it does not show sequences or produce optimal alignment.

Representation of Sequences in A Matrix

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

Dot plots are used to illustrate the concept of similarity.



Representation of Dot Plots

3. DYNAMIC PROGRAMMING

Dynamic programming is a computational method that is used to align two or more protein or nucleic acid sequences in a way that allows the computationally efficient incorporation of gaps The method is very important for sequence analysis because it provides the very best or optimal alignment between sequences. Dynamic Programming Can Provide Global Or Local Sequence Alignment.

(A) Global Alignment- A global alignment program is based on the Needleman Wunsch algorithm.

(B) Local alignment- A local alignment program on the Smith- Waterman algorithm.

Global alignment: Needleman Wunsch Algorithm: Described by Needleman and wunsch (1970), but was also proven mathematically and extended to include an improved scoring system by Smith and Waterman (1981a,b). The optimal score at each matrix position is calculated by adding the current match score to previously scored positions and subtracting gap penalties, if applicable. Each matrix position may have a positive or negative score, or 0.The Needleman- Wunsch algorithm will maximize the number of matches between the sequences along the entire length of the sequences. Gaps may also be present at the ends of sequences, in case there is extra sequence left over after the alignment. These and gaps are often, but not always, given a gap penalty.

Three steps in dynamic programming: (GLOBAL ALIGNMENT) Once you have the scoring functions set and the sequences to align, there are three steps involved in calculating the optimal scoring alignment. The three steps are as follows:

Initialization Step: In the initialization step of global alignment, each row Si,o is set to w*i. In addition, each column So,j is set to w*j. Remember, that w is the gap penalty.

Matrix Fill step: One possible solution of the matrix fill step finds the maximum global alignment score by starting in the upper left hand corner in the matrix and finding the maximal score Si,j for each position in the matrix. We can then proceed to fill in the rest of the matrix in a similar fashion. The resulting , matrix is as follows:

		G	A	A	T	T	C	A	G	T	T	A
	0	-4	-8	-12	-16	-20	-24	-28	-32	-36	-40	-44
G	-4	5	1	-3	-7	-11	-15	-19	-23	-27	-31	-35
G	-8	1	-2	- 2	-6	-10	-14	-18	-14	- 18	-22	-26
A	-12	-3	6	7	-3	-1	-5	-9	-18	-17	-21	-17
T	-16	7	2	3	12	8	4	0	-4	- 8	-12	-16
C	-20	-11	-2	-1	8	9	13	9	5	1	-3	-7
G	-24	-15	-6	5	4	5	9	10	14	10	6	2
A	-28	-19	-10	- 1	0	1	5	14	10	11	7	11

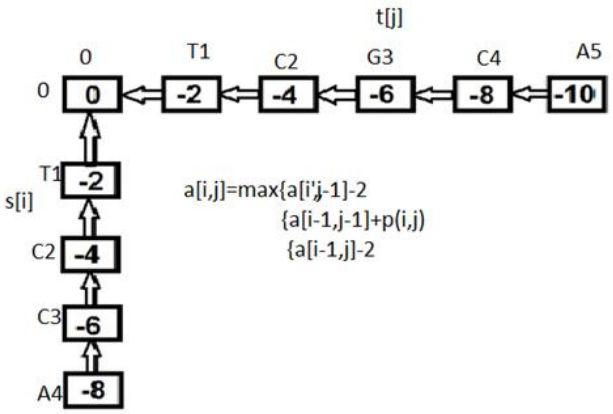
Each cell has one to three arrows indicating from which cell the maximum score was obtained. The matrix fill step is now completed.

Trace back step: After the matrix step, the maximum global alignment score for the two sequences is 11. The traceback step will obtain the actual alignment that result in the maximum score. To begin, the only possible predacessors is the diagonal match.

Local Alignment: Smith Waterman Algorithm: Local alignments are usually more meaningful than global matches because they include patterns that are conserved in the sequences. The rules for calculating scoring matrix values are slightly different, the most important differences being (1) the scoring system must include negative scores for mismatches, and (2) when a dynamic programming scoring matrix value becomes negative, that value is set to zero, which has the effect of terminating any alignment up to that point.[8]

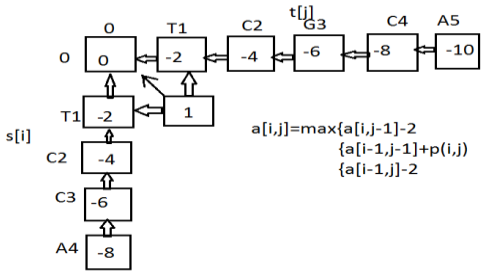
Global Alignment(GA)	**Local Alignment(LA)**
1. GA is stretched over the entire sequence length include as many matching amino acid/ nucleotide possible up to and including the sequence ends.	1. In LA the alignment stops at the ends of region of identity or strong similarity. Higher priority to find these local region
2. sequences that are quiet similar and approximately the same length are suitable candidates for GA.	2. This type of alignment favors finding conserved nucleotide patterns, DNA sequences or amino acids patterns in proteins sequences.

The first step is to calculate the alignment matrix.



Calculation of the alignment matrix

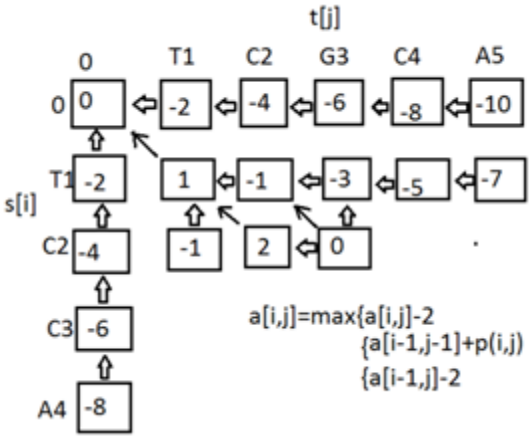
This steps involves calculation of the case in which one or more terminal gaps are added to the beginning of either sequences. This can be done by running across the top of the array and progressively adding gap penalty (Eg 12) to the score in each cell. The same in some to the left most column also. Next we apply the three scoring rules to each cell in the matrix.



Applying the scoring rules

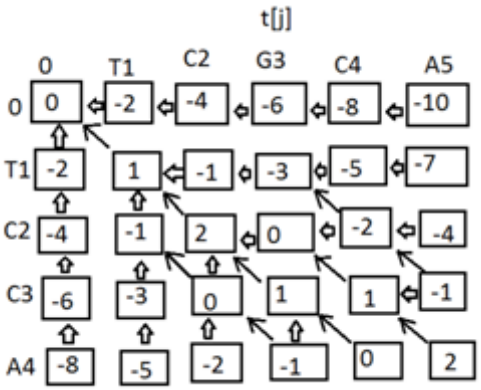
At cell a[1, 1] the score is the largest of three possible scores:
(i) a[1, 0] -2 = -2- 2 = -4
(ii) a[0, 0] + p (1, 1) = 0 + 1 = 1
(iii) a[0, 1] -2 + 2-2 + -4

Where you get =1 is s[i] equals t[j] (match) or -1 if s[i] does not equal = t[j] (mismatch). This process is shown in below figure and is repeated down the matrix until all cells are filled FFFF.



Filling up the matrix using scoring matrix

The final step shows the completed array. The arrows point from each cell to the one that is adjacent, i.e., the cell that gave the highest score or more than one adjacent cell in case of a tie. since the path to the highest scoring adjacent cell (s) is always known you can start at last position in the alignment a[m, n]. and work backwards to the beginning, a[1, 1] adding scores along a fairly limited number of alternatives paths the competed array is shown below.



Completed array

Local Alignment by Dynamic Programming

Smith Waterman dynamic programming algorithm is used for local alignment. The algorithm gives the highest-scoring local match between two sequences, Smith-Waterman is used instead of Needleman-Wunsch algorithm for matching two sequences that have a matches region that is only a fraction of their lengths, that have different lengths, that overlaps or where one sequences is fragment of the other. The rules for calculation scoring matrix values are different in this case such as: 1. The scoring system must have negative scores for mismatches.

2. When a scoring a matrix value becomes a negative, that value is set to zero, which effectively terminates the alignment up to that point.

Three steps in dynamic programming:(LOCAL ALIGNMENT)

Local alignment are produced by starting at the highest scoring positions in the scoring matrix and following a trace path from those positions up to a box that scores zero.

Initialization step: In the initialization step of local alignment, each row is set to 0. In addition, each column is set to 0.

Matrix Fill Step: One possible solution of the matrix fill step finds the maximum local alignment score by starting in the upper left hand corner in the matrix and finding the maximal score Si,j for each position in the matrix . We can proceed to fill in the rest of the matrix in a similar fashion. The resulting matrix is as follows:

		G	A	A	T	T	C	A	G	T	T	A		
	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	5	1	0	0	0	0	0	0	5	1	0	0	0
G	0	5	2	0	0	0	0	0	0	5	2	0	0	0
A	0	1	10	7	3	0	0	0	5	1	2	0	5	0
T	0	0	6	7	12	8	4	1	2	6	7	3	0	0
C	0	0	2	3	8	9	13	9	5	2	3	4	0	0
G	0	5	1	0	4	5	9	10	14	10	6	2	0	0
A	0	1	10	6	2	1	4	14	10	11	7	11	0	0

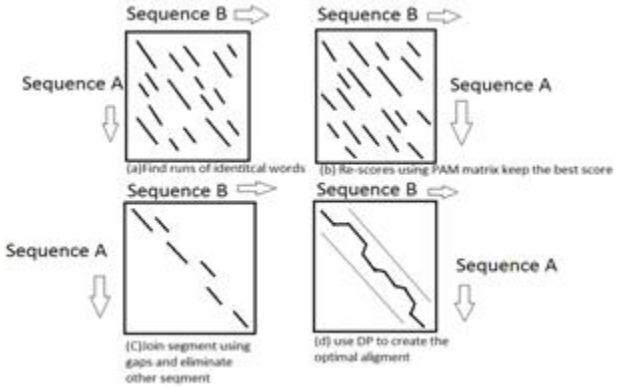
Each cell has one to three arrows indicating from which cell the maximum score was obtained. The matrix fill step is now completed.

Trace back step: After the matrix fill step, the maximum local alignment score for the two sequences is 14, which can be found by locating the highest values in the score matrix.

HEURISTICS ALGORITMS FOR SEUENCE COMPARISIONS[4]

Heuristic: method of a computer program making guesses to obtain an approximate results but much faster than possible with exhaustive searching.

FASTA: FASTA starts by making a generalization from the concept of dot plots. In a dot plot, regions of similarity between two sequences show up as diagonals. FASTA goes a step forward and calculates the sum of the dots along each.[5]

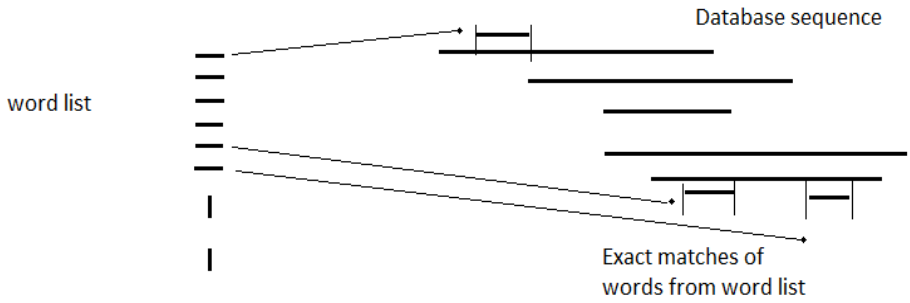


Steps of FASTA algorithm[4]

- **Find runs of identical words**. Identify regions shared by the two sequences that have the highest density of single residue identities (ktup =1) or two consecutive identities (ktup=2).
- **Re score using PAM matrix**. Keep the best score. Rescan the best regions identified in step 1 using the PAM 250 matrix. The single best score is stored as init 1 for reporting later.
- **Join segments using gaps and eliminate other segments**. Determine if gaps can be used to join the regions identified in step 2. If so determine a similarity score for the gapped alignment, which is reported as initin.
- **Use DP to create the optimal alignment**. Construct an optimal alignment of the query sequence and the library sequence (Smith- Waterman algorithm). This score is reported as the optimized score.[5]

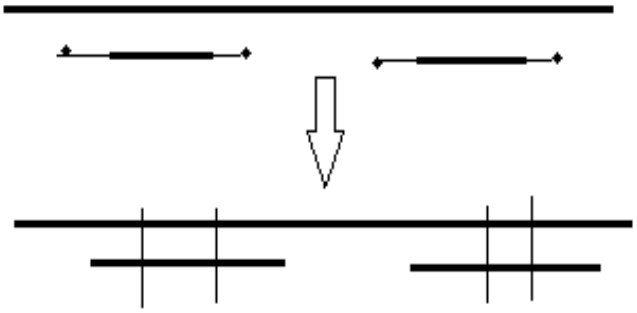
BLAST[6]

BLAST, like FASTA is a word based method. However, one major difference is that blast requires a pre-formatted search databases. BLAST algorithm is given below: 1. Find the list of high scoring words (w). BLAST takes each word from the query sequence (typically w is 3 for amino acids and 11 for nucleotides). and locates all similar words in the currents test sequences. Find the list of high scoring words w. 2. Compare the word list to the database and identify the exact matches. If similar words are found, BLAST tries to expand the alignment to the adjacent words. Without allowing or gaps.



Identity the exact matches after comparison

1. After all words are tested, a set of maximal segment pairs (MSPs) is chosen for that database sequence. Several short, non overlapping MSPs may be combined in a statistical test to create a larger, more significant matc.



Choose the MSPs

COMPARISON BETWEEN FASTA AND BLAST algorithm[6],[7]

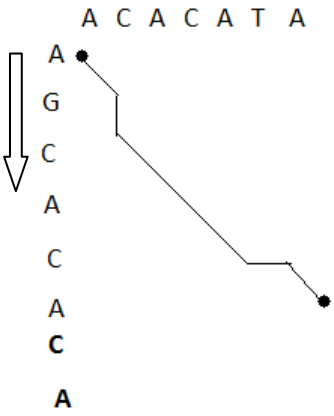
FASTA	BLASTA
FASTA is a DNA and protein sequence alignment software package.	BLAST is basic local alignment search tool
It was first described (as FASTP) by David J. Lipman and William R. Pearson in 1985.	It was developed by Altschul et al.
It searches for all matching words of length k.	It searches for most unusual or high scoring words (proteins of length 3, nucleic acids of length11).
FASTA3 calculates statistical parameters from unrelated sequences during data bases search.	BLAST2 calculates parameters for the scoring matrix and gap penalty combination and uses in database search.
FASTA3 does not remove low complexity regions.	BLAST2 removes the low complexity regions.
If a significant number of exact matches are found, FASTA uses the dynamic programming algorithm to compute optimal alignments.	Rather than requiring exact matches , it uses a scoring function to measure similarity(rather than distance).
It considers exact matches between short substrings.	It is not necessary to consider exact matches between short substrings.
It does not require a pre-foramatted search database	It requires a pre – formatted search data base
FASTA can compare DNA sequence to DNA databanks	BLAST cannot compare DNA sequence to DNA databanks

SCORING MATRICES[8],[9],[10]

Scoring matrices are used for computing alignment scores, based on observed substitution rates derived from the substitution frequencies. Let take an example of two sequences t: ACACACTA The two sequences may b rewritten in the form of a scoring matrix.

t →

		A	C	A	C	A	C	T	A
	0	1	2	3	4	5	6	7	8
A	1	0	1	2	3	4	5	6	7
G	2	1	1	2	3	4	5	6	7
C	3	2	1	2	2	3	4	5	6
A	4	3	2	1	2	2	3	4	5
C	5	4	3	2	1	2	2	3	4
A	6	5	4	3	2	1	2	3	3
C	7	6	5	4	3	2	1	2	3
A	8	7	6	5	4	3	2	2	2



Representation of a scoring matrix
(a) shows the completely filled matrix, and (b) shows the optimal path.

The most popular two scoring matrices are PAM and BLOSUM matrices. There is a third, but not so widely used method called GONNET Matrices.

1. PAM Matrices: PAM means Point accepted mutations. Margaret Dayhoff and coworkers originally proposed PAM model of evolution in the 60's. PAM-1 therefore is a scoring system for sequences in which 1% of the residues have undergone mutation and PAM -250 represents 250% mutation, i.e., an average of 2.5 accepted mutations. Two types of matrices can be constructed:[7]
1. Mutation probability matrix.
2. Relatedness odds matrix.

2. BLOSUM Matrices: Stands for blocks substitution matrix and based on PROSITE signatures **BLOSUM approaches can be summarized as follows:** The BLOSUM Matrices are best for detecting local alignment. The BLOSUM62 matrix is the best for detecting the majority of weak protein similarities and the BLOSUM45 matrix is the best for detecting log and weak alignments.[10]

REFERENCES

1. Textbook of Bioinformatics concepts, skills and applications by S.C.Rastogi, Namita Mendiratta, Parag Rastogi.
2. Textbook of Bioinformatics: Sequence and Genome analysis by David W. Mount
3. Introduction to Bioinformatics by Arthur M. Lesk.
4. https://en.wikipedia.org/wiki/Sequence_alignment

5. https://www.bioinformatics.org/wiki/Sequence_alignment
6. https://proteinstructures.com/Sequence/Sequence/sequence-alignment.html
7. https://bioinf.comav.upv.es/courses/biotech3/theory/sequence_alignment.html
8. https://www.cs.helsinki.fi/bioinformatiikka/mbi/courses/07-08/itb/slides/itb0708_slides_41-82.pdf
9. https://www.youtube.com/watch?v=8XaA_NtPt2o
10. https://www.ebi.ac.uk/Tools/psa