Study of Essential Sequence analysis tools in Bioinformatics :A BRIEF OVERVIEW

Ghafran Ali[1]* , Kanza Ashfaq[1]

[1] Institute of Biotechnology, Gulab Devi Educational Complex, Lahore,Pakistan.

Corresponding author: ghafranali994@gmail.com

Abstract

Sequence analysis program is outlined that analyzes and investigates homology between various nucleic acid or protein sequence. The dot matrix technique compares the sequences and the consensus sequence is obtained by superimposing on each other all the dot matrices. Local Alignment and Global Alignment both sequence from start to end is the best possible alignment over the entire duration between the two sequences. This method is more important to align with two closely related sequences roughly the same length. This method may not able to generate optimal results for divergent sequences and variable length sequence because between the two sequences it does not recognize very similar local region.

Keywords : Global Alignment, Local Alignment, Heuristic Algorithm , Exhaustive Algorithm

Introduction

In contemporary biology, quantitation and quantitative instruments are essential. The most biological study includes the use of some sort of mathematical, statistical or computational instruments to assist synthesize recorded data and incorporate different kinds of information in answering a specific biological query. For instance, it requires enumeration and statistical to evaluate daily laboratory experiments, such as making serial dilutions of solution or counting bacterial colonies, phage plaques, or natural environment trees and livestock. A classic instance in genetic history is Gregor Mendel and Thomas Morgan, Who were able to explore the principles of genetic inheritance by merely counting a genetic variant of crops and fruit flies.

Bioinformatics is the discipline of a computer-aided quantitative assessment of biological macromolecular data. The literature and the worldwide web contain a range of definition; some are more inclusive than others. Here,in defining bioinformatics as a union of biology and informatics,we embrace the concept suggested by Luscombe et al. in defining bioinformatics as a union of biology and computer science: bioinformatics includes technology that utilizes computers to store, retrieve, manipulate and distribute data linked to biological macromolecules such as DNA, RNA, and protein. The emphasis here is on using a personal computer because most of the function is extremely repetitive or mathematically complicated in genomic data analysis. The use of the computer is absolutely essential for the collection of data and the development of expertise in mining genomes.

Bioinformatics ‘ ultimate objective is to better comprehend a living cell and how it works at the molecular level. Bioinformatics study can produce fresh perspectives and provide a “worldwide “view of the cell by evaluating raw molecular sequence and structural information. The reason a cell’s function can be better-understood b evaluating sequence data is eventually that the flow of genetic information is dictated by the biology’s “core dogma” in which DNA is transcribed to RNA-translated Protein. Cellular function is conducted primarily by proteins whose capacity is eventually determined by their sequences.it has therefore proved to be a fruitful effort to solve functional issues using sequence and sometimes structural methods.

METHODS

Sequence alignment in parallel

As Fresh biological sequence are produced at exponential rates, sequence comparison is becoming progressively essential to draw functional and evolutionary inference from a fresh protein already present in the database.in this sort of comparison, the most basic method is sequence alignment. This is the method of comparing sequence by looking for prevalent personality patterns and creating a correspondence between associated sequence between residue and residue.

DNA and protein biological macromolecule, nucleotide bases, and amino acids building blocks from the linear sequence that determines the molecule primary structure. These molecules can be regarded as molecular fossils recording the history of millions of year of development. The molecular sequence undergoes random modification during the span, some of which are chosen during the evolution phase. As the selected sequence accumulate mutations gradually and diverge over time, traces of evolution may still remain in some portions of sequence to allow the common ancestry to be identified. The existence of developmental traces is due to fact that some of the residue performing an important function and structural roles tend to be maintained by natural selection; other residues which may be less important for structure and function tend to mutate more often.

Sequence homology/Sequence similarity

Sequence homology is a significant idea in sequence assessment. They are said to have a homology when two sequences are descended from a common evolutionary origin. A associated but distinct word is sequence resemblance, the percentage of aligned residues comparable in physiochemical characteristics such as size, load, and hydrophobicity.it is essential to differentiate sequence homology from the resemblance of associated word sequence because some scientists often confuse the two terms and use them in scientific literature interchangeably.

doi:10.20944/preprints202101.0135.v1

To be evident, sequence homology is an inference or conclusion of a common ancestral connection derived from the comparison of sequences share a sufficiently elevated degree of resemblance. On the other side, the resemblance is a direct consequence of sequence alignment observation. Using percentage, sequence similarity can be quantified; homology is a qualitative declaration. One might conclude, for instance, that two sequences share a resemblance of 40%. It's wrong to say the two sequence share a homology of 40 percent. The are homologous or non-homologous.

Sequence similarity/Sequence identity

Sequence similarity and sequence identity are another set of associated terms for sequence comparison. Sequence similarity and identification of sequences are synonymous with a sequence of nucleotides. However, the two ideas are very distinct for protein sequence. Sequence identity relates to the proportion of matches of the same residues of amino acids between two matched sequence alignment. Similarity relates to the proportion of residue aligned with comparable physicochemical features and can be more easily replaced. Sequence similarity/identity can be calculated in two ways. One includes using both sequence ' general sequence length; the other standardize by the size of the shorter sequence. The first technique utilizes the formula as follows:

$$S=[(Ls \text{ x } 2)/(La + Lb)] \text{ x } 100 \qquad \text{(Eq.1)}$$

S is the percentage sequence similarity, Ls is the number of residues aligned with similar characteristics, and La and Lb are the total lengths of each sequence.The sequence identity (1%) can be calculated in similar fashion

$$I= [(Li \text{ x } 2)/(La + Lb)] \text{ x } 100 \qquad \text{(Eq 2)}$$

Where Li is the number of aligned identical residues.

The second method of calculation is to derive the percentage of identical/similar residues over the full length of the smaller sequence using formula:

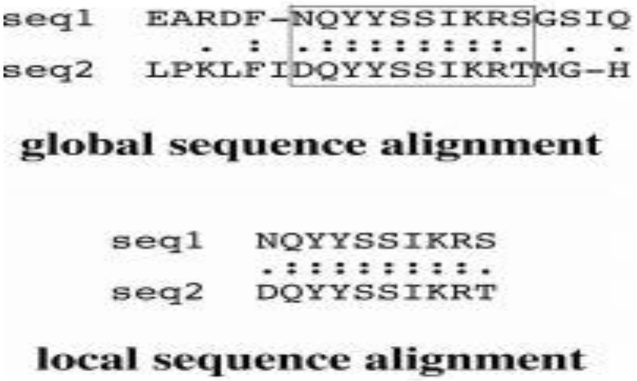
$$I(s)\%=Li(S)/La\% \qquad \text{Eq(3)}$$

Where La is the kength of shorter of the two sequenes.

Local Alignment and Global Alignment

In worldwide alignment, it is presumed that two sequences to be aligned over their entire length is usually comparable. In order to find the best possible alignment across the entire length between the two sequences, alignment is performed from beginning to end of both sequences. This technique is more relevant to align approximately the same length with two tightly associated sequences. This technique may not be able to produce ideal outcomes for divergent sequences and sequences of variable lengths because it does not acknowledge very comparable local area between the two sequences.

Local alignment, on the order hand, does not assume similarity over the entire length of the two sequences in question.it discovers only local areas with the greatest degree of resemblance between the two sequences and aligns these regions regardless of the alignment of the remaining sequence areas. This method can be used to align more divergent sequence with the objective of looking in DNA or protein for preserved sequences patterns. The two alignment sequence may have distinct lengths. This strategy is more suitable for aligning divergent biological sequence that contains only comparable modules known as domains or motifs.figure 1 Explains the distinctions between worldwide and local alignment in pairs.



```
seq1   EARDF-NQYYSSIKRSGSIQ
         . : .::::::::. . .
seq2   LPKLFIDQYYSSIKRTMG-H
```

```
seq1   NQYYSSIKRS
       .::::::::.
seq2   DQYYSSIKRT
```

figure 1 An instance of a comparison of pairwise succession illustrating the difference between worldwide and local alignment. All residues of both sequences are included in the worldwide alignment(top). A box highlights the region with the greatest resemblance. Local alignment involves only parts of the two sequences with the greatest regional resemblance.'' '' shows identical residue matches in the row between the two sequences, and ''.'' Suggests comparable residue matches.

Alignment Algorithms

Global and local alignment algorithms are fundamentally similar and differ only in the strategy of optimization used to align similar residues. Both kinds of algorithms can be based on one of the three techniques: dot matrix, dynamic programming, and word method.

Dot Matrix Method

The dot matrix technique, also known as the dot plot technique, is the most fundamental technique of sequence alignment.it is a graphical way in a two-dimensional matrix to compare two sequences. Two sequences to be compared in a dot matrix are written in matrix's horizontal and vertical axes. the comparison is made by checking for resemblance with all residues of one sequence. If a match for residues in the other sequence.if a match for residues is discovered, a dot is inside the graph. Otherwise, the positions of the matrix will remain blank. When the two sequences have significant similarity areas, many dots line up to form adjacent diagonal lines revealing the alignment of sequence.if interruptions occur in the center of a diagonal row, insertion or deletions are indicated. In the matrix parallel, diagonal lines depict repetitive sequence areas.figure 2

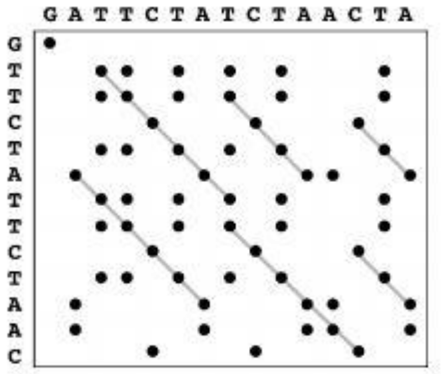


Figure2 Example to use dot plots to compare two sequences. Lines Connecting the diagonal dots show the alignment of the sequence. Diagonal lines above below the primary daignaol either sequence's inner repeats.

There is an issue when using the dot matrix technique to compare big sequence, namely the elevated noise level. Points are plotted throughout the graph in most dot plots, obscuring identification of real alignment. The issue is particularly acute for DNA Sequence because they are only four possible characters in DNA and therefore each residue has a one-in-four opportunity of matching a residue in a different sequence. In order to reduce in a different sequence. In order to reduce noise, a filtering technique must be applied instead of using a single residue to scan for similarity, Which uses a fixed-length ‘’window’’ covering a stretch of residue pairs. Windows slide across the two sequences when applying filtering to compare all possible stretches. Dots are placed only when a stretch of residues equal to the size of a window from one sequence fully matches another stretch. It has been shown that this technique is efficient in lowering noise levels. The window is also referred to as a tuple, whose size can be manipulated to plot a definite pattern of sequence match. If the chosen window size is too long, However, the alignment sensitivity will be lost.

The technique of dot matrix displays all possible matches of sequences. However, by connecting neighboring diagonals, it is often up to the user to build a complete alignment with insertion and deletions. Another restriction of this technique of visual assessment is the evaluation of the quality of the alignment of pairs. Scaling up to various alignment is hard for the technique. The following are examples of web servers that use dot plots to provide a pair sequence comparison.

Two EMBOSS programs made available online are **Dotmatcher**(bioweb.pasteur.fr/seqanal/interfaces/dottup.html). Dotmatcher aligns and shows dot plots in FASTA format with two input sequence(DNA or proteins). Use a defined length window

doi:10.20944/preprints202101.0135.v1

and a scoring system. Diagonal lines are plotted only over the place of the window if the resemblance exceeds a certain limit. Diagonal lines are drawn only if there are accurate matches of designated length phrase.

Matrix plot (www.cbs.dtu.dk/services/MatrixPlot/) is a more advanced protein and nucleic acid sequence alignment matrix plot program.The user has the choice to add from recognized three-dimensional protein or nucleic acid structure such as sequence logo profiles and distance matrices. The program utilizes colored grids instead of using dots and rows to show alignment or other user-defined data.

Dynamic programming Method

Dynamic programming is a technique of determining optimal alignment by combining two sequences between the two sequences for all possible pairs of characters. The dot matrix technique is essentially comparable in that is also generates a two-dimensional alignment grid. However, By turning a dot matrix into a scoring to account for matches and discrepancies between sequence, it discovers alignment in a more quantitative manner. The best alignment can be correctly achieved by looking for the set of greatest results in a matrix.

Dynamic programming operates by first building a two -dimensional matrix whose axes are the two comparable sequences. The matching of residues is based on a specific scoring matrix. One row at a moment, The results are calculated. This begins with the first line of series used to scan the full duration of the other sequence, followed by the second-row scanning. It calculates the corresponding results. The second-row scanning requires the results already acquired in first round into consideration. The highest score is placed in an intermediate matrix's lower correct corner (figure 3). The process is iterated until all cells are filled with values. The results are therefore collected along the diagonal from the top left corner to the bottom correct corner. The next step is to discover the route that represents the ideal alignment once the scores have been accumulated in the matrix. This is performed by tracing back in reverse order through the matrix from the matrix's reduced right corner toward the matrix's origin in the upper left corner. The best way to match is the one with the highest complete score. (figure 3) If two or more routes achieve the same maximum score, one is arbitrarily selected to represent the best alignment. At a certain stage, the route may also shift horizontally or vertically, corresponding to the implementation of gap or entry or deletion for one of the two sequences.

Gap Penalties

Optimum sequence alignment often involves the application of gaps representing insertions and deletions. Because insertion and deletion are relatively rare in nature evolutionary process compared to substitutions, the introduction of gaps should be made computationally more difficult, reflecting the rarity of evolutionary insertion and deletion events. However, assigning Pena o appeal values may be more or less arbitrary, as there is no evolutionary theory to determine a precise cost of inserting and deleting. If the penalty values are set too low, gaps may be become too numerous to allow elevated similarity scores to match even non-related sequences. If the penalty values are set too high, gaps may be become too hard to appear and it is not possible to achieve sensible alignment, which is also unrealistic. A set of penalty values have been created through empirical research for globular proteins that seem to fit most aims of alignment. In most alignment programs, they are usually implemented as default values.

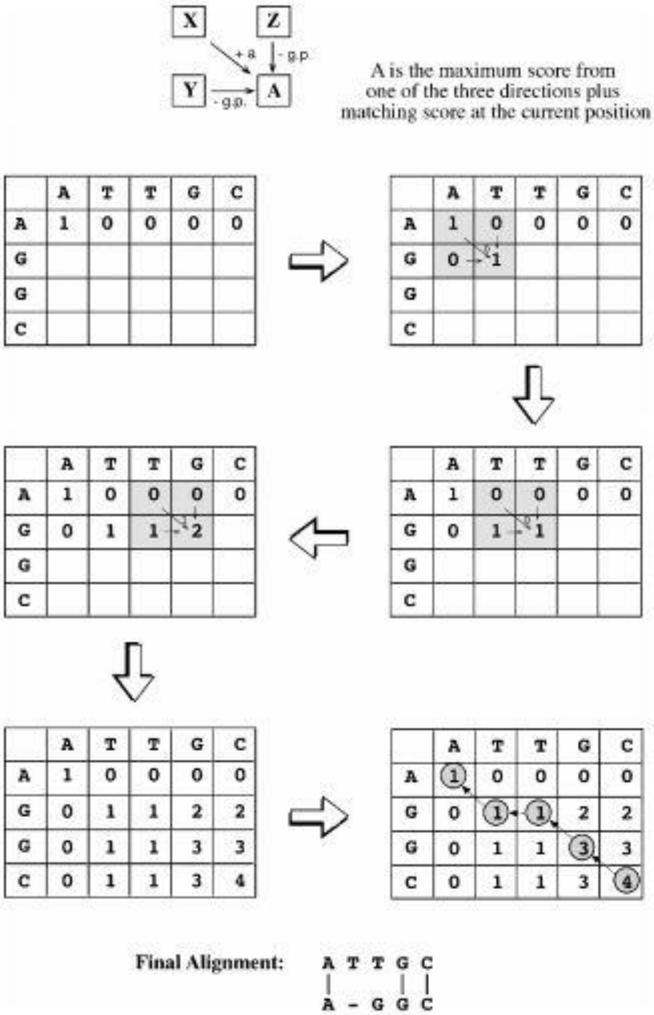

X
Z
Y
A
+a
-g.p.
-g.p.
A is the maximum score from one of the three directions plus matching score at the current position
A T T G C
A G G C
Final Alignment:
A T T G C
A - G G C

figure 3 Example of two sequences pairs of alignment using dynamic programming. The score for bottom right square (A) of a 2/2 matrix is the maximum square score from one of the other three neighboring squares (X, Y, and Z) plus and minus the exact single residue match score (a) for the bottom right corner and the gap penalty (g.p.) For the two brief sequences, a matrix is set up. A easy scoring scheme in which a score of 1, a mismatch of a score of 0, and a gap penalty are allocated to an identical game is -1. The results in the matrix are filled from top to bottom one row at a moment and one cell at a moment. According to this rule, the highest results are filled to the bottom correct corner of the sub matrix(gray boxes).When all the cells are filled with the highest complete score is determined through a trace-back operation. A penalty is implemented when a route passes horizontally or vertically.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2021 doi:10.20944/preprints202101.0135.v1

Dynamic programming For local alignment

The amount of divergence between the two sequences to be aligned is not readily understood is regular sequence alignment. There may also be unequal sequence length s of the two sequences. In such instances, it may be more important to identify regional sequence resemblance than to find a match that involves all residues. The smith- waterman algorithm is the first implementation of dynamic programming in local alignment. Positive results are allocated in this algorithm to match residues and zeros to match mismatches. There is no use of adverse results.

In dynamic programming, a comparable tracing-back method is used. However, along the main diagonal, the alignment route may start and end internally. It begins with the lowest scoring position and continues diagonally up to the left until it reaches a zero cell. If needed, gaps will be placed.in this situation, the punishment for the affine gap is often used. Occasionally, several sections with the highest results are achieved that are optimally aligned. The final outcome, as in the worldwide alignment, is affected by the selection of scoring schemes used(to be outlined next). The objective of local alignment is obtained the greatest alignment score locally, which for a full length may be at the cost of general score possible. This strategy may be appropriate to align divergent sequence or sequence with various domains that may originate differently. The most frequently used web servers for pair alignment apply the local alignment approach.SIM, LALIGN, and SSEARCH.

SSEARCH (http://pir.georgetown.edu/pirwww/search/pairwise.html) is a straightforward web-based program using the smith-waterman sequence alignment. Only one alignment with the highest score is a provided.No choice is available to score matrices a penalty gap scores.

Dynamic Programming For Global Alignment

The Needleman-Wunsch algorithm is the classical worldwide pairwise algorithm using dynamic programming. An idea alignment is achieved throughout the lengths of two sequences in this algorithm. To achieve the highest total score, It must extend from the beginning to the end of both sequences. In other words, from the bottom correct corner of the matrix, the alignment route has to go the top left corner. The drawback of concentrating on having a maximum score for sequence alignment in full length is the danger of missing the highest local resemblance. This approach can only be used to align two tightly associated sequence of the same duration. The strategy does not generate optimum alignment for divergent sequences or sequences with distinct domain structures. One of the few worldwide pairwise alignment web servers is GAP.

GAP (http:/bioinformatics.iastate.edu/aat/align/align.html) is a worldwide alignment program based on a pair of websites. It aligns two sequences so that comparable sequences of unequal lengths can be aligned without penalizing terminal gaps. Such gaps are handled with a steady punishment in order to be to insert length gaps in the alignment. In aligning cDNA with exons in genomic DNA containing the same gene, This characteristics his helpful.

Data similarity Searching

This method includes submitting a query sequence and comparing the query sequence in pairs with all individual sequences in a database. Thus, Looking for database resemblance is a big scale pair-by-pair alignment. This sort of search in one of the most efficient ways of assigning putative features to new sequences. There are distinctive requirements for sequence database search algorithms to be implemented. Sensitivity is the first criterion, referring to the capacity to find as many right hits as possible. It is measured by the extent to which sequence members of the same family are included. In the database search practice, these right hits are regarded as " real positive." The second criterion is selectivity, referring to be the capacity to exclude wrong hits. These inaccurate hits are erroneously recognized unrelated sequences in the search for databases and are deemed "false positives." The third criterion is velocity, Which is the time it takes for database search outcomes to be obtained. Depending on the database size, Velocity may be a major problem at times.

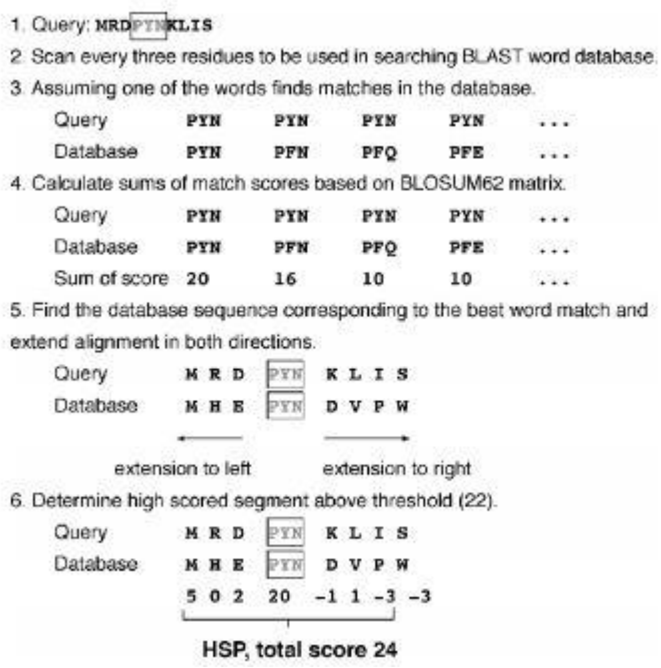
Heuristic Database Searching

Two significant heuristic algorithms are currently available for performing database search: BLAST and FAST. These techniques are not guaranteed to discover the ideal alignment or real homologs, But 50-100 times quicker than dynamic programming. The enhanced computational velocity comes at a moderate cost of search sensitivity and specificity that work molecular biologists can readily tolerate. By recognizing comparable sequence section, both programs can provide a fairly excellent indication of sequence resemblance.

BLAST LOCAL ALIGNMENT SEARCH TOOL (BLAST)

NCBI's Stephen Altschul developed the BLAST program in 1990 and has since become one of the programs in 1990 and has since become one of the most popular sequence analysis programs.BLAST utilizes heuristics to align a series of queries with all database sequences. The goal is to discover ungapped sections of high scoring among associated sequences. The presence of such sections above a specified limit shows an over random resemblance in pairs, which helps to discriminate associated sequence from the unrelated sequence in the database.

By the following steps, BLASt conducts sequence alignment. Creating a list of phrases from the query sequence is the first step. Each term typically consists of three protein sequence residues and eleven DNA sequence of queries. It's also called seeding this step. The second stage is to search for the occurrence of these phrases in a sequence database that contains the corresponding phrases. A specified substitution matrix scored the matching of phrases. if it above a limit, a term is regarded as a match. The fourth stage includes pair alignment by expanding the alignment score with the same substitution matrix from the phrases in both directions while counting. The expansion goes on until the alignment score falls below a mismatches limit (the drop point is twenty-two for proteins and twenty for DNA). High scoring segment pair(HSP; see work in figure 4) is the resulting adjacent aligned segment pair without gaps. The best rated HSPs are provided as the final document in the initial BLAST version. They also referred to as peak pair scoring.



1. Query: MRDPYNKLIS

2. Scan every three residues to be used in searching BLAST word database.

3. Assuming one of the words finds matches in the database.

```
Query      PYN   PYN   PYN   PYN   ...
Database   PYN   PFN   PFQ   PFE   ...
```

4. Calculate sums of match scores based on BLOSUM62 matrix.

```
Query          PYN   PYN   PYN   PYN   ...
Database       PYN   PFN   PFQ   PFE   ...
Sum of score   20    16    10    10    ...
```

5. Find the database sequence corresponding to the best word match and extend alignment in both directions.

```
Query      M R D  PYN  K L I S
Database   M H E  PYN  D V P W
```

6. Determine high scored segment above threshold (22).

```
Query      M R D  PYN  K L I S
Database   M H E  PYN  D V P W
           5 0 2  20   -1 1 -3 -3
```

(Figure 4) BLAST procedure illustration using a hypothetical query sequence that matches a hypothetical database sequence. The term match instance will be illustrated in the box.

VARIANTS

BLAST is a BLASTN, BLASTP, BLASTX, TBLASTN and TBLASTX family of programs.with a nucleotide sequence database, BLASTN is queried nucleotide sequences . For searching against a protein sequence database, BLASTP utilizes protein sequences as queries.BLASTX utilizes nucleotide sequences as queries and translated protein sequences in all six reading frames that are used to query a database of a protein sequence.TBLASTN queries protein sequences with the sequences translated in all six reading frames into nucleotide sequence database.TBLASTX utilizes nucleotide sequence to search for a nucleotide sequence database that has all the sequences translated into six frames in all six frames.

BLAST Output Format

The output of the BLAST includes a graphical overview box, a matching list, and an alignment text description. The graphical overview panel includes colored horizontal bars that enable the number of database hits and degrees of hit resemblance to be quickly identified. The horizontal bar color coding corresponds to the sequence hits sequence similarity ranking(red: most associated; green and blue: mildly black: unrelated).The bar duration reflects the sequence alignment spans relative to the sequence of queries.

FASTA

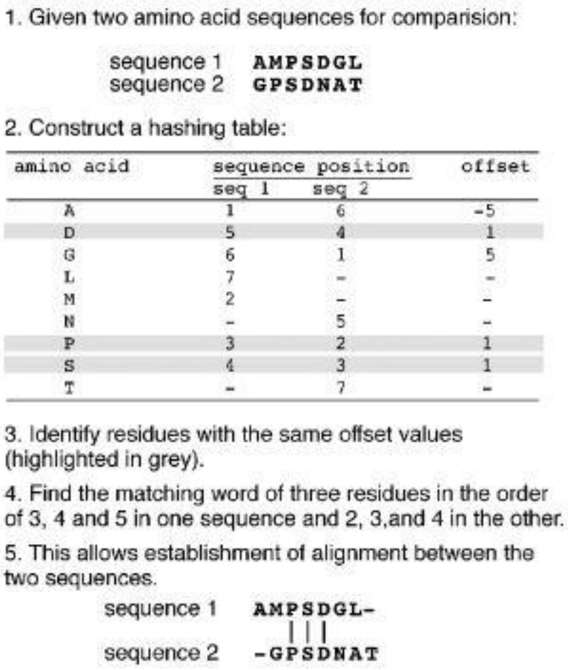
In reality, FASTA (FAST ALL, www.ebi.ac.uk/fasta33/) was the first similarity database search instrument created before BLAST was Created. For a brief stretch of identical residues with of K, FASTA utilizes a ‘’hashing’’ approach to find matches. The residue string is referred to as tuples or ktups, which are equal to BLAST words, But are usually shorter than phrases. A ktup typically consists of two protein sequence residues and six DNA sequence residues.

The first stage in FASTA alignment is to use the hashing approach to define ktups between two sequences. This approach operates by creating a lookup table showing each ktup’s position for the two sequences being considered. By subtracting the position of the first sequence, the positional difference of each term between the two sequences is acquired and expressed as the offset.

The ktups with the same offset values are then connected to show a contiguous identical sequence region in a two-dimensional matrix that corresponds to a stretch of diagonal. (figure 5)

The second step is to narrow between the two sequences the elevated resemblance areas. Normally, in the hashing phase, many diagonals can be recognized between the two sequences. The top ten areas with the greatest diagonal density are recognized as areas with the elevated resemblance. In these areas, the diagonals are scored using a matrix of replacement. There are chosen and joined neighboring high –scoring sections along the same diagonal to form a single alignment. This step enables gaps between the diagonals to be introduced while implementing penalties for gaps.

The gapped alignment score is again calculated.In step 3, the gapped alignment is further refined using the Smith-waterman algorithm to create a final alignment.The last step is to conduct a final alignment statistical assessment as in BLAST, which generates the E-value.



1. Given two amino acid sequences for comparision:

```
sequence 1   AMPSDGL
sequence 2   GPSDNAT
```

2. Construct a hashing table:

amino acid	sequence position		offset
	seq 1	seq 2	
A	1	6	-5
D	5	4	1
G	6	1	5
L	7	-	-
M	2	-	-
N	-	5	-
P	3	2	1
S	4	3	1
T	-	7	-

3. Identify residues with the same offset values (highlighted in grey).

4. Find the matching word of three residues in the order of 3, 4 and 5 in one sequence and 2, 3,and 4 in the other.

5. This allows establishment of alignment between the two sequences.

```
sequence 1   AMPSDGL-
               |||
sequence 2   -GPSDNAT
```

(Figure 5) The ktup recognition method using FASTA ‘s hashing approach.Identical offset values between the two sequence residues enable ktups to be formed.

Multiple Sequence Alignment

Multiple sequences is a natural expansion of pairwise alignment, which is to align various associated sequences for ideal sequence matching. Related sequences are recognized by looking for resemblance in the database. Since the process generates multiple matching sequence pairs, It is often necessary to covert the numerous pair alignments into a single alignment, Which arranges sequences in such a way that equivalent positions are matched across all sequences.

Multiple sequence alignment has a d distinctive benefit because it shows more biological data than many alignments in pairs can. It enables, for instance, to identify conserved sequence patterns and motifs in the entire sequence family, which are not apparent to detect by comparing only two sequences. Multiple alignments of proteins can identify many preserved and functionally critical residues of amino acids. Multiple sequence alignment is also an important precondition for conducting sequence family phylogenetic analysis and prediction of secondary and tertiary protein structures. Multiple sequence alignment also has applications based on the various associated sequence in the design of degenerate polymerase chain reaction (PCR) primers.

Exhaustive Algorithms

Multiple sequence alignment uses exhaustive and heuristic methods. The exhaustive technique of alignment includes the simultaneous examination of all possible aligned positions. Similar to dynamic programming in pair alignment, involving the use of a two-dimensional matrix to search for optimal alignment, to use dynamic programming for multiple sequence alignment, additional dimensions are required to take into account all possible methods of matching sequences. This implies setting up a multidimensional matrix of search. For example, to account for all possible alignment results, a three-dimensional matrix is needed for sequence. Back-tracking is implemented to discover the best-scored route represents the idea alignment through the three-dimensional matrix. To align the N sequence, it is necessary to fill an N-dimensional matrix with alignment results. As the amount of computational time and memory space required increase exponentially with the number of sequences, using the method for a large data set is computationally prohibitive. Full dynamic programming is therefore restricted to tiny data sets of less than 10 brief sequences. For the same reason, there is few multiple alignment programs openly accessible using this ''brute force'' strategy. A program called DCA, which uses some exhaustive components.

DCA (Divide-and-Conquer Alignment, http:/bibiserv.techfak.uni-bielefld.de/dca/) is a semi-exhaustive web-based program because some computing steps are limited into narrower parts each of sequences. The breakpoints are determined based on the sequences regional resemblance. If the segments are not sufficiently brief, there will be further divisions. When the sequence lengths achieve a predefined limit, dynamic programming is implemented to align each subset. The resulting brief in head to tail to alignment are combined head to tail to produced multiple alignments of the entire sequence duration. This algorithm offers an alternative to use a heuristic operation (fastDCA) to select optimal cutting points so that a higher amount of sequence can be handled more quickly.it conducts worldwide alignment and needs comparable lengths and domain structures of the input sequences. The program is still highly computationally intensive despite the use of heuristics and can manage only datasets with a very restricted amount of sequence.

Results

Dot Matcher :-

Dotmatcher generates a dotplot from two input sequences .The dotplot is an initiative graphical representatives of regions of similarity between two sequences.

Procedure :-

1. Dot Matcher tool is opened by **EMBOSS** at the URL,

(https://www.ebi.ac.uk/Tools/seqstats/emboss_dotmatcher/)

2. Select any portion of the sequence from NCIBI, (Protein).

3. Select these portions of sequence from two different sources, i.e from NCBI a nd copied in Fasta format, (Human and mus-musculus).

4. Copy these sequences and paste in the sequence window of dotmatcher.

5. Set the threshhold 23.

6. After setting run the dotmatcher.

Results :-

- After aligning these two sequences by **Needleman-Wunsch** algorithm.

- A diagonal line is observed on the graph show's that both of these sequences are found to be similar.(See figure no. 6)

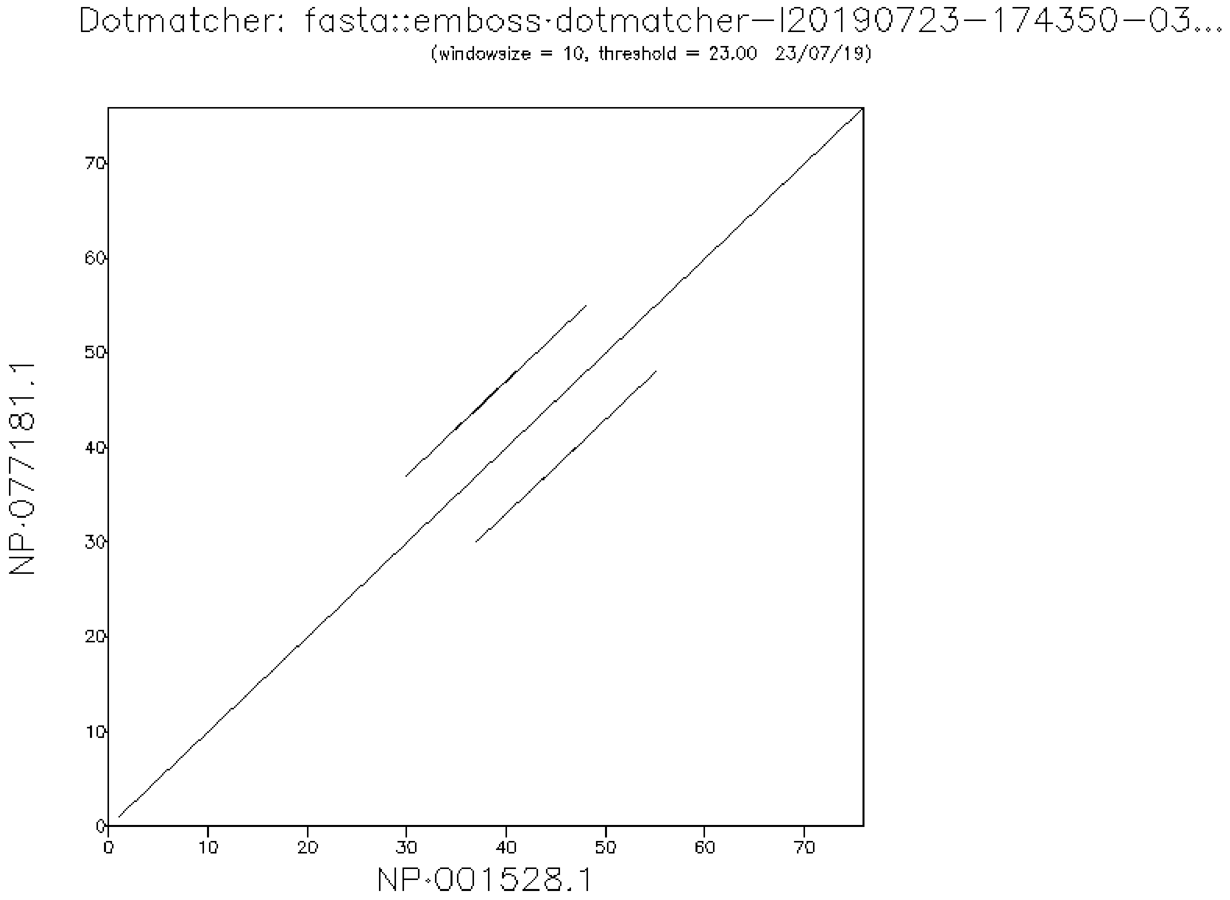


Figure no.6 Heat Shock factor –binding protein 1 [Homo sapiens] and heat shock factor binding protein 1 [Mus musculus] that alignment show by dot matcher as daignoal form in their certain thresholder 23 and windowsize 10.

Global Sequence Alignment :-

For protein sequence alignment, we use **Needleman-Wunsch** algorithm.

Protein:-

- **Heat shock factor-Binding protein**

Sources:- Heat Shock factor –binding protein 1 [Homo sapiens]

Heat shock factor binding protein 1 [Mus musculus]

Procedure:-

The global alignment can be performed by using online tool, EMBOSS Needle. It can be performed in two steps.

Step 1:-

1. To download the data and get access to the tools , go to the stimulater tab.
2. Get access to the tool **EMBOSS Needle.**
3. Copy and paste the Fasta format sequences.
4. One can choose the file through file option and can upload the sequence file.
5. Similarly copy and paste the second sequence file. See figure no.7

Step 2:-

1. **EMBOSS Needle** is predefined with the scoring matrices, i.e **BLOSUM-62** for protein sequences.

2. The gap open and gap extended penalities can be changed by user defined values.

3. Submit the e-mail address in the checkbox so that the user is notified with the result through e-mail.

4. After that submit the job by clicking the submit button and the results are displayed within few seconds.

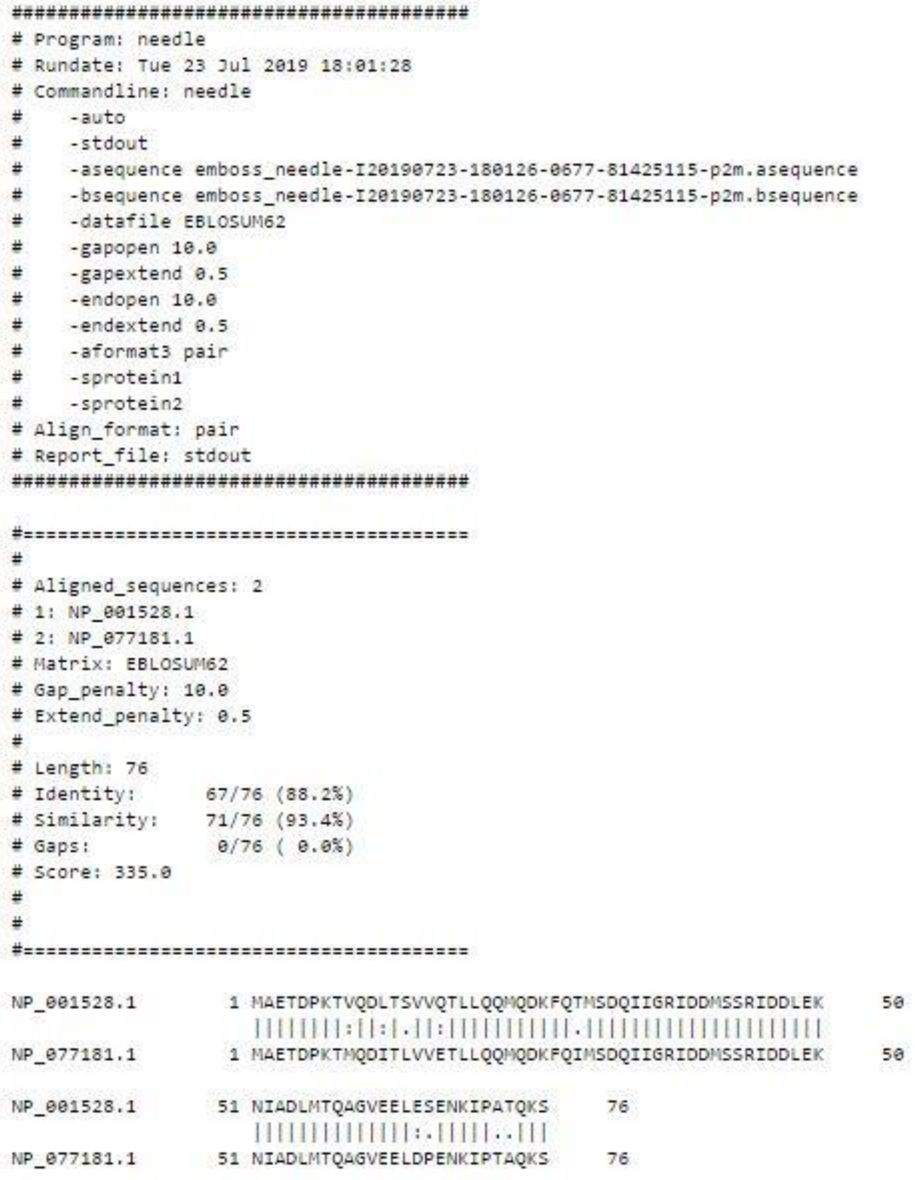
Results:-

The result page is comprises of three tabs , namely;

- Alignment
- Submission Details
- Submit Another Job

Identification:-

- Gaps are represented by, “ - ”.
- Match is represented by, “ 1 ”.
- Mismatch is represented by, “ . ”.



```
########################################
# Program: needle
# Rundate: Tue 23 Jul 2019 18:01:28
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20190723-180126-0677-81425115-p2m.asequence
#    -bsequence emboss_needle-I20190723-180126-0677-81425115-p2m.bsequence
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: NP_001528.1
# 2: NP_077181.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 76
# Identity:      67/76 (88.2%)
# Similarity:    71/76 (93.4%)
# Gaps:           0/76 ( 0.0%)
# Score: 335.0
#
#
#=======================================

NP_001528.1        1 MAETDPKTVQDLTSVVQTLLQQMQDKFQTMSDQIIGRIDDMSSRIDDLEK     50
                     ||||||||:||:|.||:|||||||||||.|||||||||||||||||||||
NP_077181.1        1 MAETDPKTMQDITLVVETLLQQMQDKFQIMSDQIIGRIDDMSSRIDDLEK     50

NP_001528.1       51 NIADLMTQAGVEELESENKIPATQKS     76
                     ||||||||||||||:.|||||..|||
NP_077181.1       51 NIADLMTQAGVEELDPENKIPTAQKS     76
```

Figure no. 7 Aligment detail which occurring between Human and Mus Musclus heat shock protein by Global alignment.

Local Sequence Alignment:-

Local sequence alignment is one of the most important method of dynamic programming. In it **Waterman-Smith** Algorithm is used. It is used to provide the information about local regions, i.e conserved domains and motifs. It is used to compare two divergent sequences.

- **Protein:- Heat shock factor-Binding protein**

Source:- Heat Shock factor –binding protein 1 [Homo sapiens]

Heat shock factor binding protein 1 [Mus musculus]

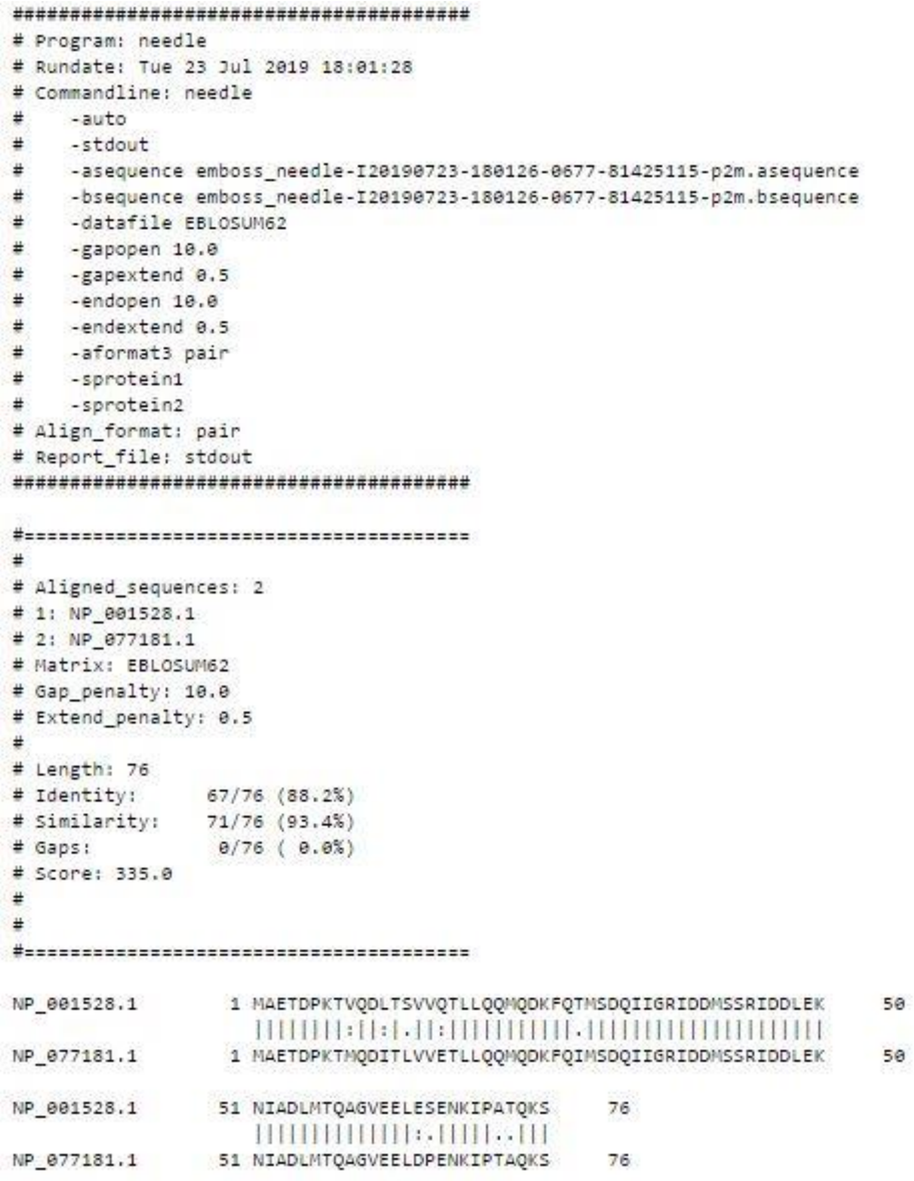
How to find local alignment:-

- Copy the Fasta format sequence.
- Enter the URL, (www.ebi.ac.uk/tools/psa/emboss-matcher).
- Enter or paste the sequence in the sequence window.
- Then set the default matrix as BLOSUM 62.
- Adjust the gap open penality as 14 and gap extension penality as 04.
- Run the job by clicking the submit button.see figure 8
-

Results:-

The result page is same as global alignment and comprises of three tabs;

- Alignment.
- Submission Details.
- Submit Another Job.



```
########################################
# Program: needle
# Rundate: Tue 23 Jul 2019 18:01:28
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20190723-180126-0677-81425115-p2m.asequence
#    -bsequence emboss_needle-I20190723-180126-0677-81425115-p2m.bsequence
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: NP_001528.1
# 2: NP_077181.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 76
# Identity:      67/76 (88.2%)
# Similarity:    71/76 (93.4%)
# Gaps:           0/76 ( 0.0%)
# Score: 335.0
#
#
#=======================================

NP_001528.1        1 MAETDPKTVQDLTSVVQTLLQQMQDKFQTMSDQIIGRIDDMSSRIDDLEK     50
                     ||||||||:||:|.||:|||||||||||.|||||||||||||||||||||
NP_077181.1        1 MAETDPKTMQDITLVVETLLQQMQDKFQIMSDQIIGRIDDMSSRIDDLEK     50

NP_001528.1       51 NIADLMTQAGVEELESENKIPATQKS     76
                     ||||||||||||||:.|||||..|||
NP_077181.1       51 NIADLMTQAGVEELDPENKIPTAQKS     76
```

Figure no.8 Alignment detail Human and Mus Musclus Heat Shock Protein by Local Alignment.

Multiple Sequence Alignment:-

Multiple sequence alignment (MSA) is used to align three or more related sequences so to achieve maximum matching between them. The goal of MSA is to arrange a set of sequences in such a way that as many characters from each sequence are matched according to some scoring formation.

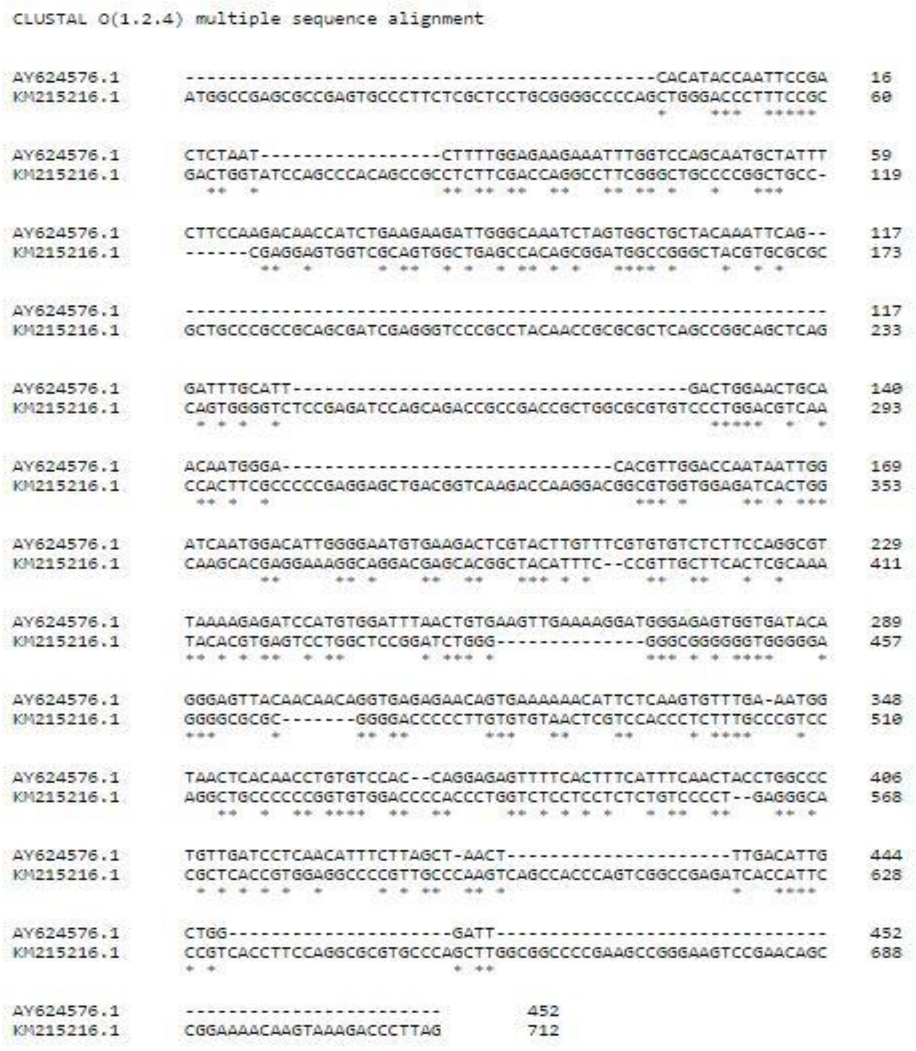
Procedure:-

- There are a number of online tools available for MSA but here we have used **Clustal Omega** and it works on progressive alignment construction.

- Go to; (http/www.ebi.ac.uk/tools/msa/clustaly).
- A page was opened to select the type of data, (protein,DNA and RNA) , sequences can be entered in both file and Fasta format.
- Then select output form, here we used amino acid sequence of **hspB1 of , Cucmis melo , Bubalus bubalis, ,** After entering sequences and selecting parameters , submit the job by clicking the submit button or option.

Results:-

- The results of the job can be viewed as figure no. 9
- The results can be downloaded as a file by clicking download alignment file button.



```
CLUSTAL O(1.2.4) multiple sequence alignment

AY624576.1      --------------------------------------------CACATACCAATTCCGA    16
KM215216.1      ATGGCCGAGCGCCGAGTGCCCTTCTCGCTCCTGCGGGGCCCCAGCTGGGACCCTTTCCGC    60
                                                            *    ***  *****

AY624576.1      CTCTAAT-----------------CTTTTGGAGAAGAAATTTGGTCCAGCAATGCTATTT    59
KM215216.1      GACTGGTATCCAGCCCACAGCCGCCTCTTCGACCAGGCCTTCGGGCTGCCCCGGCTGCC-    119
                  **  *                 ** ** **  **   ** ** *   *   ***

AY624576.1      CTTCCAAGACAACCATCTGAAGAAGATTGGGCAAATCTAGTGGCTGCTACAAATTCAG--    117
KM215216.1      ------CGAGGAGTGGTCGCAGTGGCTGAGCCACAGCGGATGGCCGGGCTACGTGCGCGC    173
                       **  *      * **  * *  * ** * *   **** *    *  * *

AY624576.1      ------------------------------------------------------------    117
KM215216.1      GCTGCCCGCCGCAGCGATCGAGGGTCCCGCCTACAACCGCGCGCTCAGCCGGCAGCTCAG    233

AY624576.1      GATTTGCATT-------------------------------------GACTGGAACTGCA    140
KM215216.1      CAGTGGGGTCTCCGAGATCCAGCAGACCGCCGACCGCTGGCGCGTGTCCCTGGACGTCAA    293
                 * * *  *                                        *****  *  *

AY624576.1      ACAATGGGA-------------------------------CACGTTGGACCAATAATTGG    169
KM215216.1      CCACTTCGCCCCCGAGGAGCTGACGGTCAAGACCAAGGACGGCGTGGTGGAGATCACTGG    353
                 ** *  *                                  *** *     ** * ***

AY624576.1      ATCAATGGACATTGGGGAATGTGAAGACTCGTACTTGTTTCGTGTGTCTCTTCCAGGCGT    229
KM215216.1      CAAGCACGAGGAAAGGCAGGACGAGCACGGCTACATTTC--CCGTTGCTTCACTCGCAAA    411
                       **     ** *    **  **   *** * *     **  **   *  *

AY624576.1      TAAAAGAGATCCATGTGGATTTAACTGTGAAGTTGAAAAGGATGGGAGAGTGGTGATACA    289
KM215216.1      TACACGTGAGTCCTGGCTCCGGATCTGGG--------------GGGCGGGGGGTGGGGGA    457
                ** * * **  * **       * *** *              *** * * ****    *

AY624576.1      GGGAGTTACAACAACAGGTGAGAGAACAGTGAAAAAACATTCTCAAGTGTTTGA-AATGG    348
KM215216.1      GGGGCGCGC-------GGGGACCCCCTTGTGTGTAACTCGTCCACCCTCTTTGCCCGTCC    510
                ***     *       ** **       ***   **    **     * ****    *

AY624576.1      TAACTCACAACCTGTGTCCAC--CAGGAGAGTTTTCACTTTCATTTCAACTACCTGGCCC    406
KM215216.1      AGGCTGCCCCCCGGTGTGGACCCCACCCTGGTCTCCTCCTCTCTGTCCCCT--GAGGGCA    568
                   **  *  ** ****  **  **     ** * * * *   * **  **    ** *

AY624576.1      TGTTGATCCTCAACATTTCTTAGCT-AACT---------------------TTGACATTG    444
KM215216.1      CGCTCACCGTGGAGGCCCCGTTGCCCAAGTCAGCCACCCAGTCGGCCGAGATCACCATTC    628
                 * * * * *  *     * * **  ** *                     *   ****

AY624576.1      CTGG---------------------GATT-------------------------------    452
KM215216.1      CCGTCACCTTCCAGGCGCGTGCCCAGCTTGGCGGCCCCGAAGCCGGGAAGTCCGAACAGC    688
                * *                      * **

AY624576.1      ------------------------        452
KM215216.1      CGGAAAACAAGTAAAGACCCTTAG        712
```

Conclusion

Alignment derscribing fresh protein alignment techniques will be releasesd.While Many of these methods are based on the same fundamental principles, implementation are based on the same fundamental principles, implementation details can have drastic impacts on efficiency, both in terms of precision and velocity.Due to complex size of sequences and the search space,Arranging molecular sequences within an alignment to find similarities and differences between them is not an easy task.Genetic algorithum is used as a real alterative to multiple sequence alignment issue due to the copacity to mange complicated scale issue.

References.

1. Batzoglou, S. 2005. The many faces of sequence alignment.Brief.Bioinformatics 6:6-22
2. Brenner,S E, Chothia, C., and Hubbard, T. J 1998.Assessing sequence comparsion methods with reliable structureally identified distant evolutionary relationship.Proc. Natl Acad. Sci.U S A 95:6073-8
3. Chao,K-m., Pearson, W.R. and Miller, w.1992.Aligning two sequence within a specified diagonal band. Comput. Appl.Biosci. 8:481-7
4. Henikoff,S., and Henikoff, and Henikoff, J. G 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U SA 89:109-19
5. Huang X. 1994.On global sequence alignment .Comput. Appl. Biosci. 10:227-35
6. Pagni, M., and jongeneel, V. 2001.Making sense of score statistics for sequence alignments. Brief.Bioinformatics 2:51-67
7. Pearson, W. R. 1996 Effective protein sequence comparison.Method Enzymol.266:227-58.
8. Rost, B. 1999.Twilight Zone of protein sequence alignments.Protein Eng. 12:85-94.
9. States, D. J., Gish, W., and Altschul, S.F 1991. Improved sensitivity of nucleic acid database searches using application-specific scoring matrices.Methods 3:66-70
10. Vingron, M., and waterman M.S. 1994. Sequence alignment and penalty scores. J . Mol. Biol. 235:1-12.
11. Altschul, S. F., Boguski, M. S., Gish, Wish., and Wottom, J. C. 1994. Issues in searching molecular sequence database Nat. Genet. 6:119- 29.
12. Altschul, S. F., Madden T. L., Schaffer, A. A., Zhang, J., and Zhang., Miller, W., and Lipman,D. J. 1997. Gapped BLAST and PSI-BLAST : A new generation of protein database serach programs. Nuclic acids Res. 25:3389-402
13. Chen, Z. 2003. Assessing sequence comparison methods with the average precision criterion.Bioinformatics 19:2456-60
14. Mullan, L. J., and Williams, G, W.2002 BLAST and go? Brief. Bioinform.3:200-2
15. Sansom, C., 2000. Database searching with DNA and protein sequences: An introduction. Brief Bioinform. 1:22-32

16. Spang R., and Vingron, M. 1998. Statistics of large-scale sequence searching. Bioinformatics 14:279-84
17. Attwood, T. K., and Miller, C. J. 2002 Progress in bioinformatics and the importance of being earnest.Biotechnol. Annu. Rev. 8:1-54
18. Golding G.B. 2003 DNA and the revolution of molecular evolution, computational biology and bioinformatics Genome 46:930-5
19. Goodman,N. 2002. Biological data becomes computer literature :New advances in bioinformatics.Curr. Opin. Biotechnol. 3:68-71
20. Hagen. J. B. 2000. The origin of bioinformatics. Nat Rev. Genetics 1:231-6.
21. FASTA www.ebi.ac.uk/fasta33/
22. SSEARCH http://pir.georgetown.edu/pirwww/search/pairwise.html)
23. GAP http:/bioinformatics.iastate.edu/aat/align/align.html
24. DOT Matcherhttps://www.ebi.ac.uk/Tools/seqstats/emboss_dotmatcher/
25. Ali, G.; Ashfaq, K. Brief Overview. Role of Computation Biology & Bioinformatics in Drug Design. Preprints 2020, 2020070478 (doi: 10.20944/preprints202007.0478.v2).
26. Ali, G.; Sharif, M.; Jabeen, A.; Kabira, M.U.; Ashfaq, K. HBV, HCV, HIV & TB Prevalence in Injection Drug Users in Major Cities of Punjab, Pakistan- a Survey-Based Research Report. Preprints 2021, 2021010070 (doi: 10.20944/preprints202101.0070.v1).