Date: 26-10-21

WEBLEM 10

Introduction to Pairwise Alignment

Pairwise Alignment:

Sequence comparison lies at the heart of bioinformatics analysis. It is an important first step toward structural and functional analysis of newly determined sequences. As new biological sequences are being generated at exponential rates, sequence comparison is becoming increasingly important to draw functional and evolutionary inference of a new protein with proteins already existing in the database. The most fundamental process in this type of comparison is sequence alignment. This is the process by which sequences are compared by searching for common character patterns and establishing residue-residue correspondence among related sequences. Pairwise sequence alignment is the process of aligning two sequences and is the basis of database similarity searching and multiple sequence alignment. This chapter introduces the basics of pairwise alignment.

Sequence Similarty vs Sequence Identity:

Another set of related terms for sequence comparison are sequence similarity and sequence identity. Sequence similarity and sequence identity are synonymous for nuclotide sequences. For protein sequences, however, the two concepts are very differenct. In a protein sequence alignment, sequence identitty refers to the percentage of matches of the same animo acid residues between two aligned sequences. Similarity refers to the percentage of aligned residues that have similar physicochemical characteristics and can be more readily substituted for each other.

Needleman-Wunsch Algorithm:

Needleman-Wunsch Algorithm developed by Saul B. Needleman and Christian D. Wunsch in 1970. It was designed to compare biological sequences and was one of the first applications of dynamic programming to the biological sequence comparison. This algorithm is usually used for global alignment of two sequences (nucleotide or amino acids). For the purpose of explanation, first summarizing the algorithm in five steps and then I will move forwards with examples and explanations.

1) Consider all the possible pairs of residues from two sequences, the best way is to generate a 2D matrix of two sequences. We will need 2 such matrices one for sequence and one for scores.
2) Initialize the score matrix. Scoring matrices are used to determine the relative score made by matching two characters in a sequence alignment. There are many flavors of scoring matrices for amino acid sequences, nucleotide sequences, and codon sequences, and each is derived from the alignment of "known" homologous sequences. These alignments are then used to determine the likelihood of one character being at the same position in the sequence as another character.
3) Gap penalty: Usually there are high chances of insertions and deletions (indels) in biological sequences but one large indels is more likely rather than multiple small indels in a given sequences. In order to tackle this issue we give two kind of penalities; Gap opening penalty (relatively higher) and gap extension penalty (relatively lower)
4) Calculate scores and fill the traceback matrix
5) Deduce the alignment from the traceback matrix

Smith-Waterman alignment algorithm:

Over a decade after the initial publication of the Needleman-Wunsch algorithm, a modification was made to allow for local alignments (Smith and Waterman, 1981). Today, the Smith-Waterman alignment algorithm is the one used by the Basic Local Alignment Search Tool (BLAST) which is the most cited resource in biomedical literature. In this adaptation, 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.

Recall that for global alignments the value at any point is:

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

Mi-1,j + gap. Mi,j-1 + gap)

However for local alignments the score calculation becomes:

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.

Date: 23-10-21

WEBLEM 10/A

(URL:https://www.ebi.ac.uk/Tools/psa/)

Aim:

To Study the query “Actin Sequence” in Needleman-Wunsch algorithm

Introduction:

Needleman-Wunsch Algorithm developed by Saul B. Needleman and Christian D. Wunsch in 1970. It was designed to compare biological sequences and was one of the first applications of dynamic programming to the biological sequence comparison. This algorithm is usually used for global alignment of two sequences (nucleotide or amino acids). For the purpose of explanation, first summarizing the algorithm in five steps and then I will move forwards with examples and explanations.

Actin is the most abundant protein in most eukaryotic cells. It is highly conserved and participates in more protein-protein interactions than any known protein. These properties, along with its ability to transition between monomeric (G-actin) and filamentous (F-actin) states under the control of nucleotide hydrolysis, ions, and a large number of actin-binding proteins, make actin a critical player in many cellular functions, ranging from cell motility and the maintenance of cell shape and polarity to the regulation of transcription. Moreover, the interaction of filamentous actin with myosin forms the basis of muscle contraction.

Methodology:

1. Open the homepage of Pair-Wise Alignment.
2. After that open the Needle-Wusch algorithm.
3. Enter the protein sequence.
4. Open the result.
5. Interpret the result.

Observation:

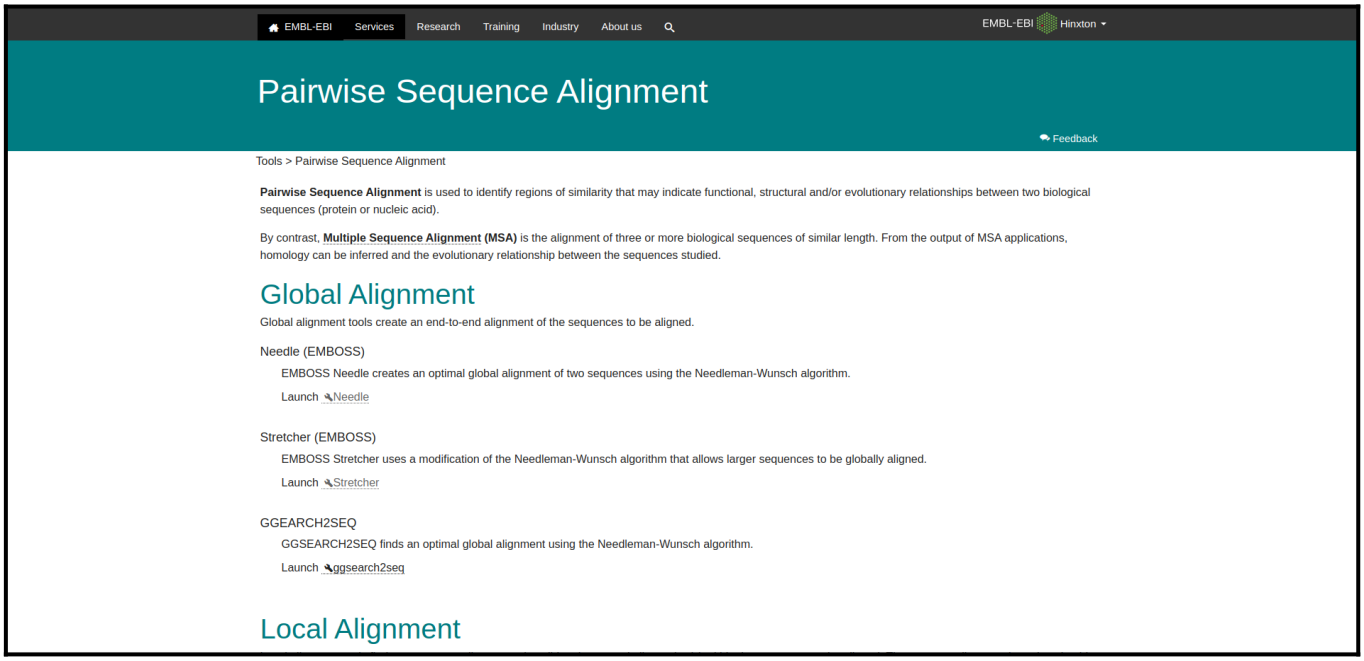

Fig1: Homepage of pairwise Sequence Alignment

```
>sp|P07830|ACT1_DICDI Major actin OS=Dictyostelium discoideum OX=44689 GN=act1 PE=1 SV=2
MDGEDVQALVIDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHTGVMVGMGQKDSYVGDEAQ
SKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPEEHPVLLTEAPLNPKANREKM
TQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVSHTVPIYEGYALPHAILRLD
LAGRDLTDYMMKILTERGYSFTTTAEREIVRDIKEKLAYVALDFEAEMQTAASSSALEKS
YELPDGQVITIGNERFRCPEALFQPSFLGMESAGIHETTYNSIMKCDVDIRKDLYGNVVL
SGGTTMFPGIADRMNKELTALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISK
EEYDESGPSIVHRKCF

>sp|P63268|ACTH_MOUSE Actin, gamma-enteric smooth muscle OS=Mus musculus OX=10090 GN=Actg2 PE=1 SV=1
MCEEETTALVCDNGSGLCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQ
SKRGILTLKYPIEHGIITNWDDMEKIWHHSFYNELRVAPEEHPTLLTEAPLNPKANREKM
TQIMFETFNVPAMYVAIQAVLSLYASGRTTGIVLDSGDGVTHNVPIYEGYALPHAIMRLD
LAGRDLTDYLMKILTERGYSFVTTAEREIVRDIKEKLCYVALDFENEMATAASSSSLEKS
YELPDGQVITIGNERFRCPETLFQPSFIGMESAGIHETTYNSIMKCDIDIRKDLYANNVL
SGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISK
PEYDEAGPSIVHRKCF
```

Fig2: Two FASTA sequence for actin from uniprot

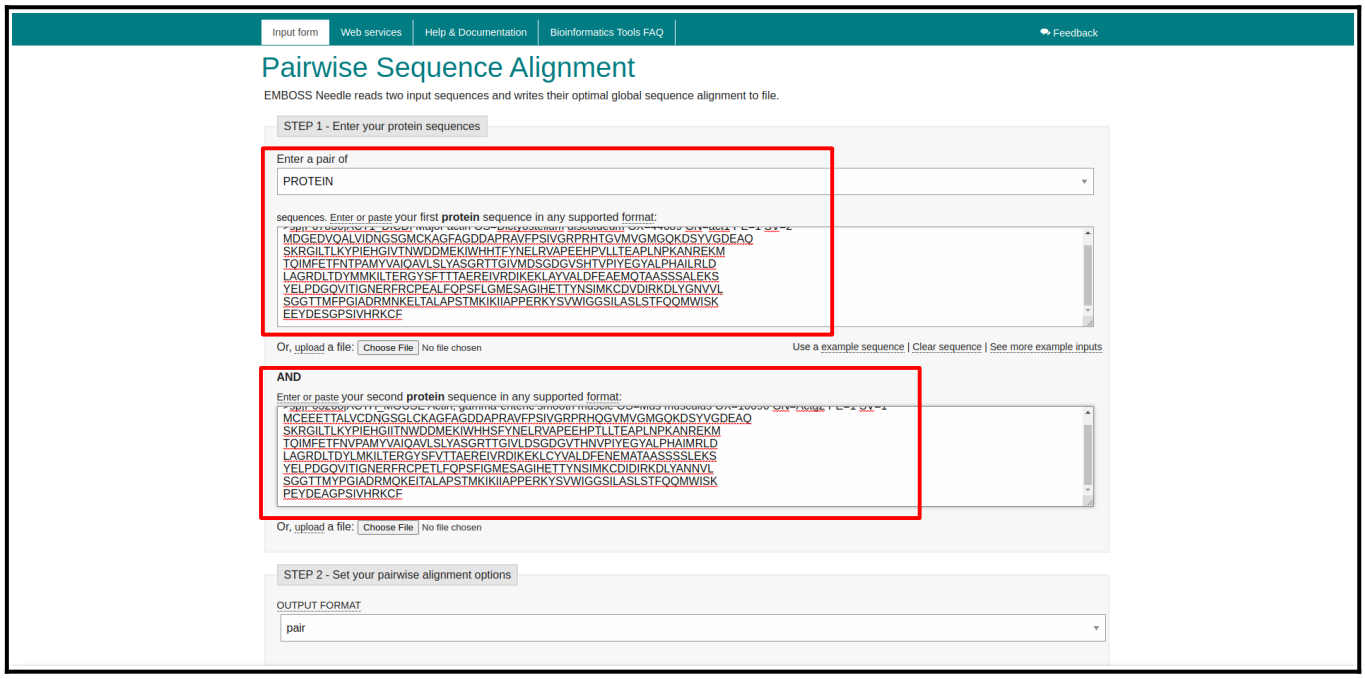

Fig3: Needleman-Wunsch algorith search bar with two actin sequences

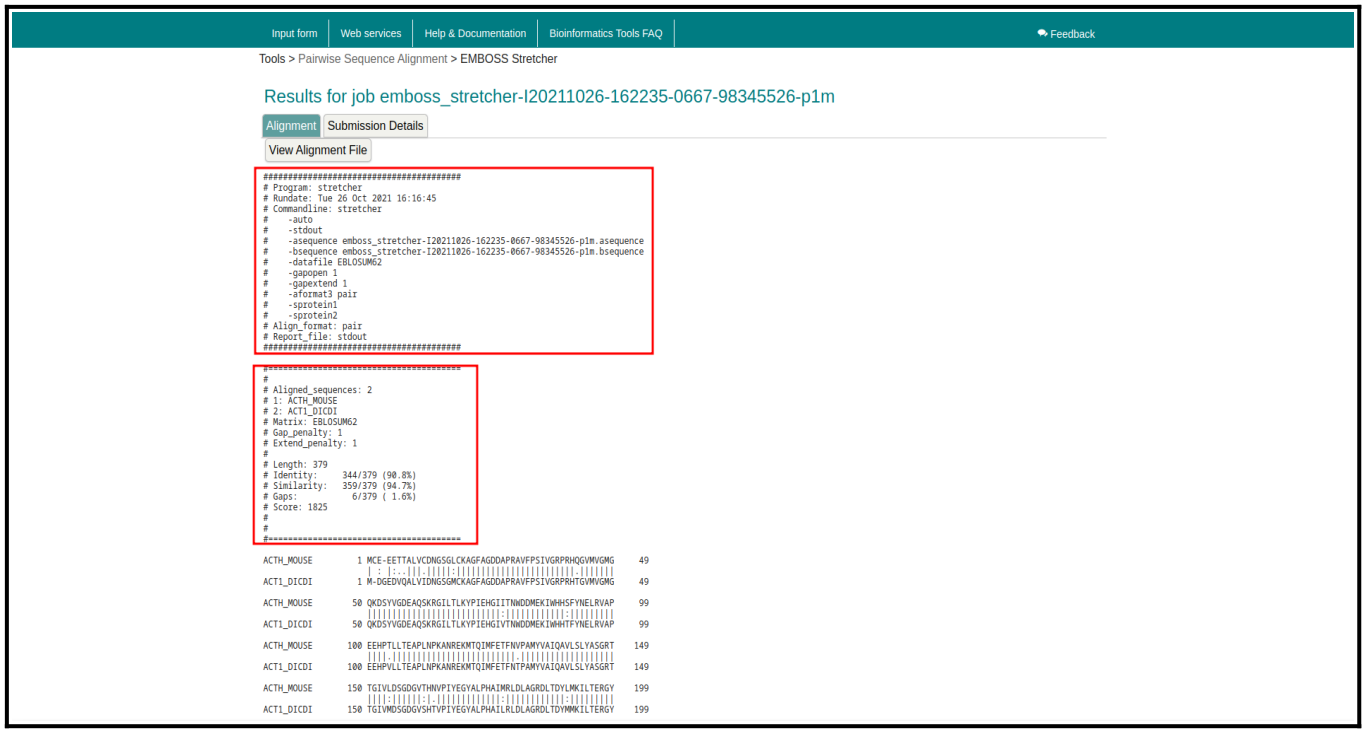

Fig4: Result page for actin Two Sequences

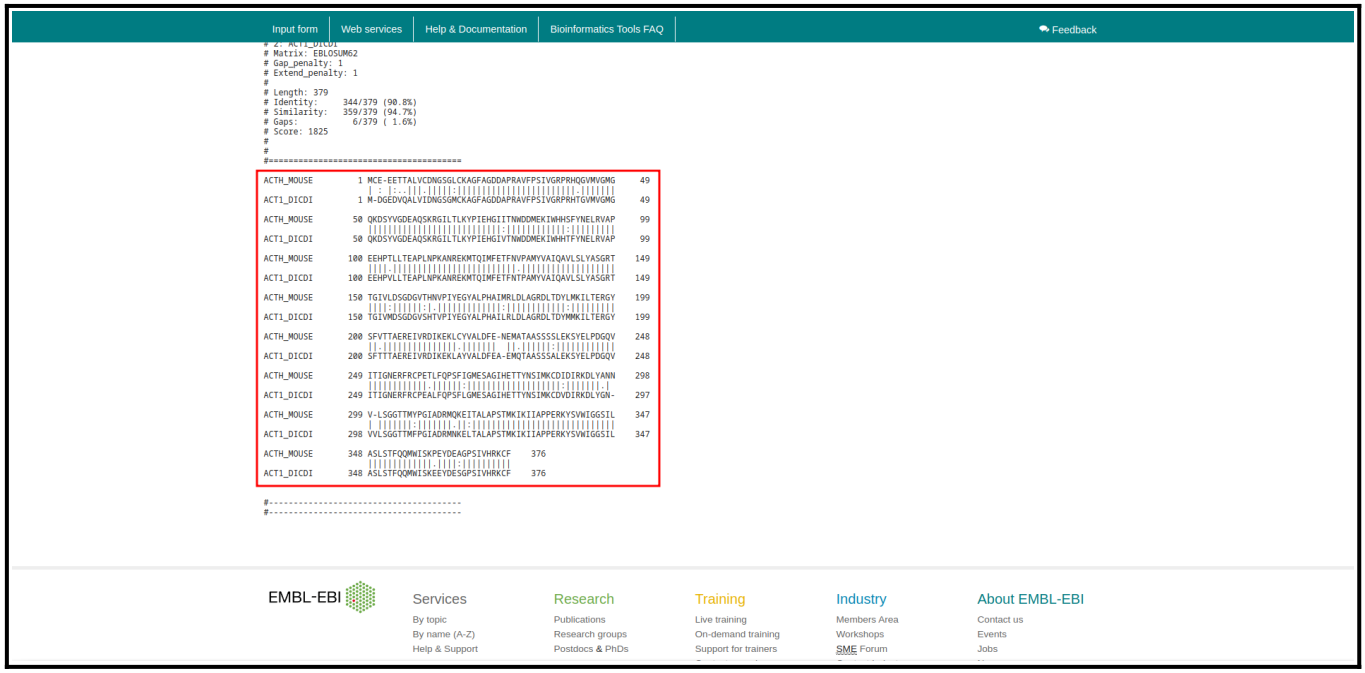

Fig5: Two Sequences Alignment for Actin

Results:

In Needleman-Wunsch algorithm for actin sequences it showing score 1825. It has length 379, Identity 344/379 (90.8%), Similarity 359/379 (94.7%), Gaps 6/379 (1.6%). At matrix EBLOSUM62 with gap penalty 1 and extend penalty 0.5

Conclusion:

Needleman-Wunsch algorithm in bioinformatics to align protein or nucleotide sequences. It was one of the first applications of dynamic programming to compare biological sequences. The Needleman-Wunsch algorithm is stil widely used fro optimal global alignment, particularly when the quelity of the global alignment is of the utmost importance. The algorithm assigns a score to every possible alignment, and the purpose of the alogrithm is to find all possible alignment having the highest score.

References:

1. Xiong, J. (2006). Essential Bioinformatics (1st ed.). Cambridge University Press.
2. (n.d.). Retrieved from https://www.uniprot.org/uniprot/P63268.fasta
3. (n.d.). Retrieved from https://www.uniprot.org/uniprot/P07830.fasta
4. Dominguez, R., & Holmes, K. C. (2011). Actin structure and function. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3130349/
5. Embl-Ebi. (n.d.). Pairwise Sequence Alignment. Retrieved from https://www.ebi.ac.uk/Tools/psa/
6. Embl-Ebi. (n.d.). EMBOSS Needle. Retrieved from https://www.ebi.ac.uk/Tools/psa/emboss_needle/
7. Embl-Ebi. (n.d.). EMBOSS Stretcher. Retrieved from https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=emboss_stretcher-I20211026-162235-0667-98345526-p1m
8. Embl-Ebi. (n.d.). EMBOSS Stretcher. Retrieved from https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=emboss_stretcher-I20211026-162235-0667-98345526-p1m

Date: 23-10-21

WEBLEM 10/B

(URL: https://www.ebi.ac.uk/Tools/psa/)

Aim:

To study the query “Actin” in Smith-Waterman alignment algorithm

Introduction:

Over a decade after the initial publication of the Needleman-Wunsch algorithm, a modification was made to allow for local alignments (Smith and Waterman, 1981). Today, the Smith-Waterman alignment algorithm is the one used by the Basic Local Alignment Search Tool (BLAST) which is the most cited resource in biomedical literature. In this adaptation, 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.

Actin is the most abundant protein in most eukaryotic cells. It is highly conserved and participates in more protein-protein interactions than any known protein. These properties, along with its ability to transition between monomeric (G-actin) and filamentous (F-actin) states under the control of nucleotide hydrolysis, ions, and a large number of actin-binding proteins, make actin a critical player in many cellular functions, ranging from cell motility and the maintenance of cell shape and polarity to the regulation of transcription. Moreover, the interaction of filamentous actin with myosin forms the basis of muscle contraction.

Methodology:

1. Open the homepage of Pair-Wise Alignment.
2. After that open the Smith-Waterman alignment algorithm.
3. Enter the protein sequence.
4. Open the result.
5. Interpret the result.

Observation:

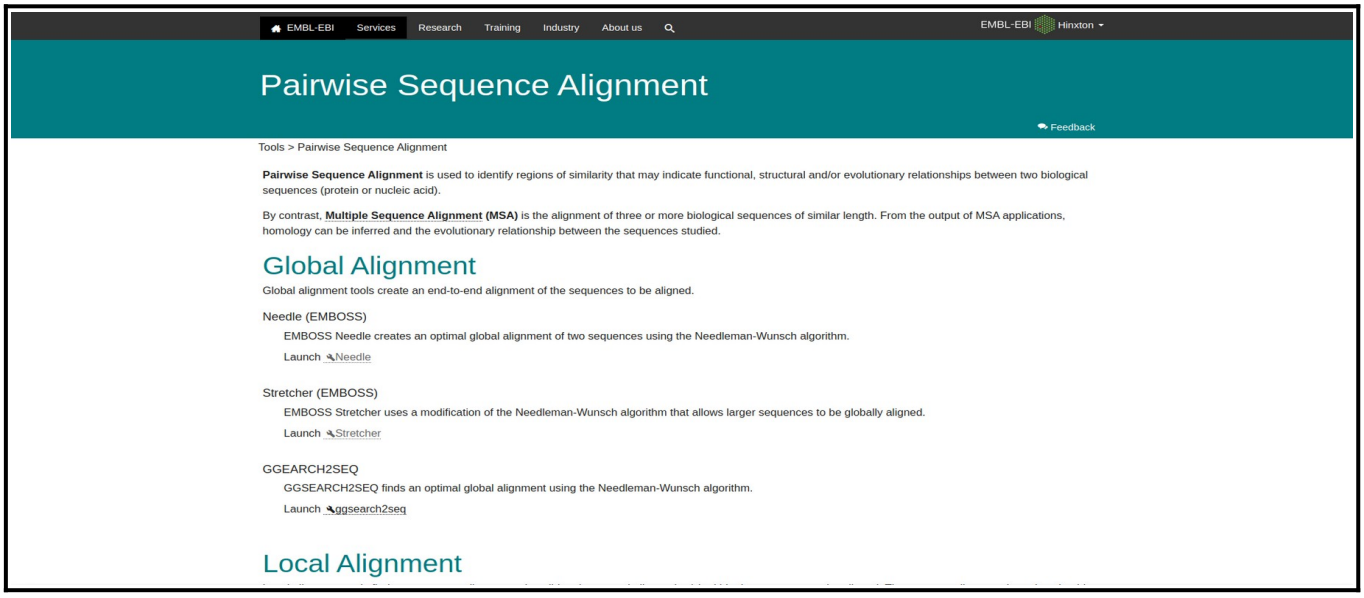

Fig1. Homepage of pairwise Sequence alignement

```
>sp|P07830|ACT1_DICDI Major actin OS=Dictyostelium discoideum OX=44689 GN=act1 PE=1 SV=2
MDGEDVQALVIDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHTGVMVGMGQKDSYVGDEAQ
SKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPEEHPVLLTEAPLNPKANREKM
TQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVSHTVPIYEGYALPHAILRLD
LAGRDLTDYMMKILTERGYSFTTTAEREIVRDIKEKLAYVALDFEAEMQTAASSSALEKS
YELPDGQVITIGNERFRCPEALFQPSFLGMESAGIHETTYNSIMKCDVDIRKDLYGNVVL
SGGTTMFPGIADRMNKELTALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISK
EEYDESGPSIVHRKCF

>sp|P63268|ACTH_MOUSE Actin, gamma-enteric smooth muscle OS=Mus musculus OX=10090 GN=Actg2 PE=1 SV=1
MCEEETTALVCDNGSGLCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQ
SKRGILTLKYPIEHGIITNWDDMEKIWHHSFYNELRVAPEEHPTLLTEAPLNPKANREKM
TQIMFETFNVPAMYVAIQAVLSLYASGRTTGIVLDSGDGVTHNVPIYEGYALPHAIMRLD
LAGRDLTDYLMKILTERGYSFVTTAEREIVRDIKEKLCYVALDFENEMATAASSSSLEKS
YELPDGQVITIGNERFRCPETLFQPSFIGMESAGIHETTYNSIMKCDIDIRKDLYANNVL
SGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISK
PEYDEAGPSIVHRKCF
```

Fig2. Two FASTA sequence for actin from uniprot

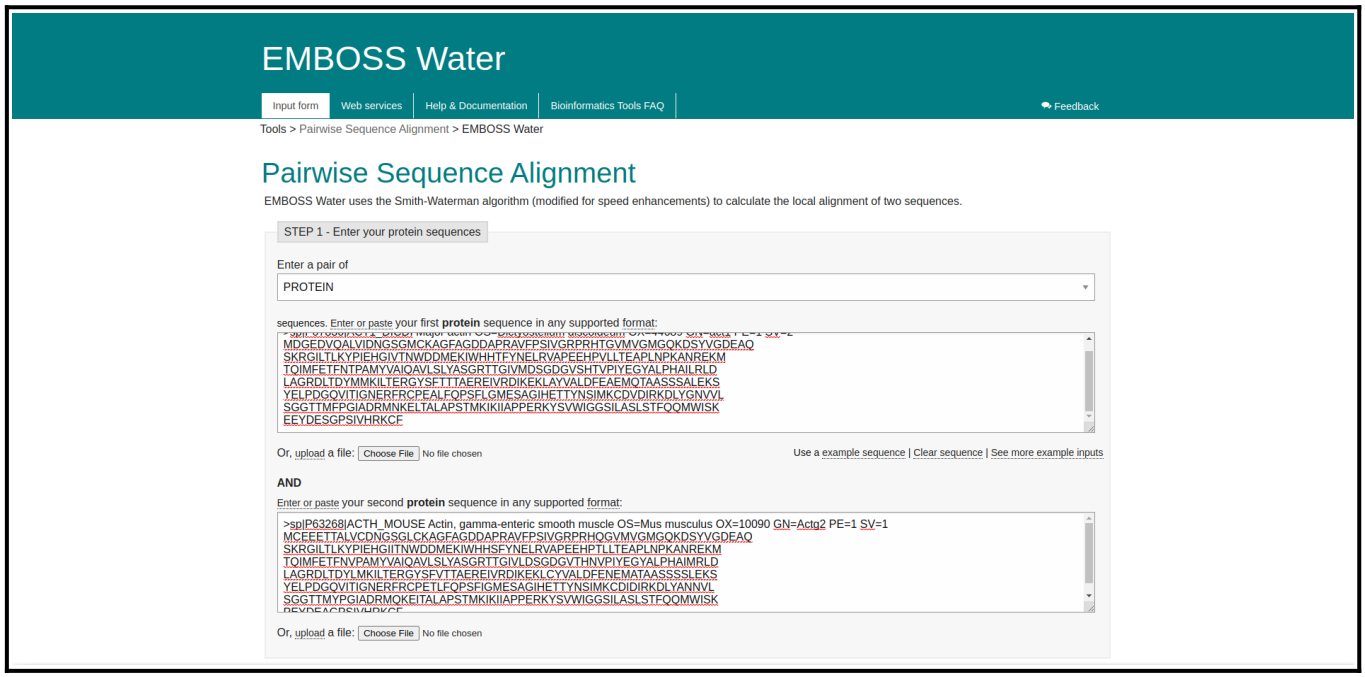

Fig.3 Smith-Waterman alignment algorithm search bar with 2 actin sequences

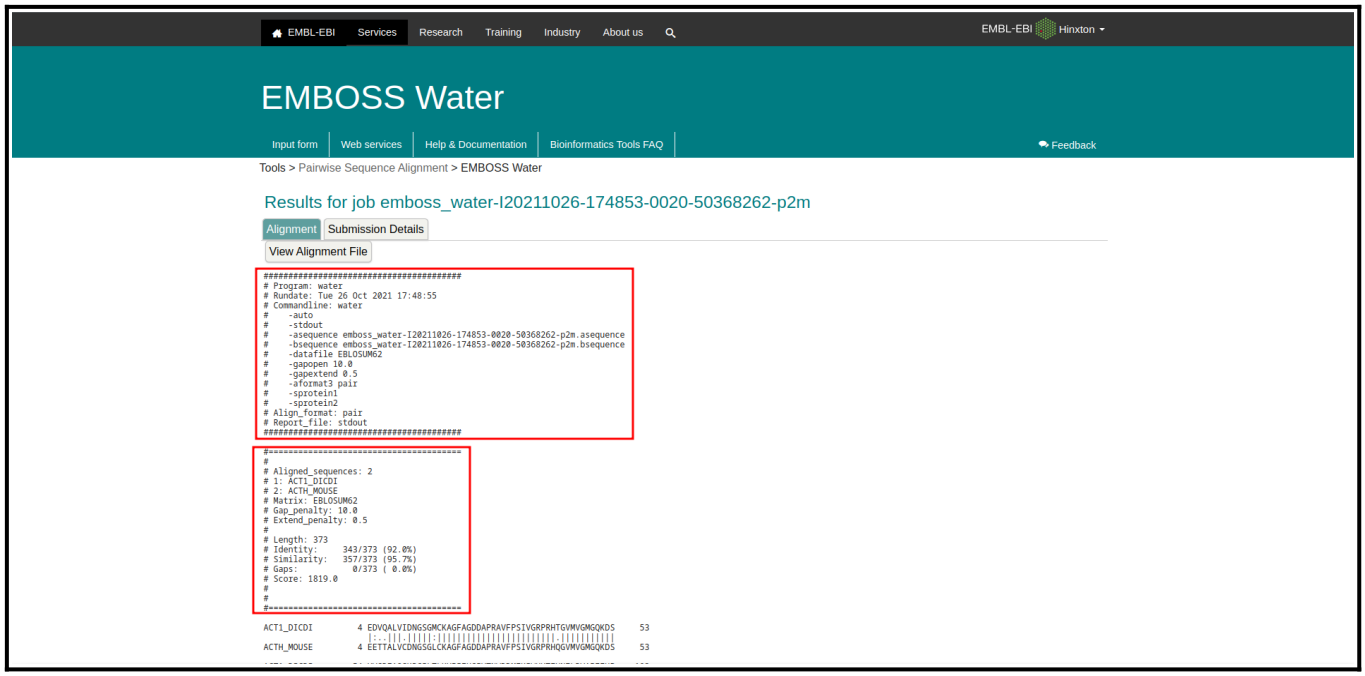

Fig4. Result page for 2 actin sequences

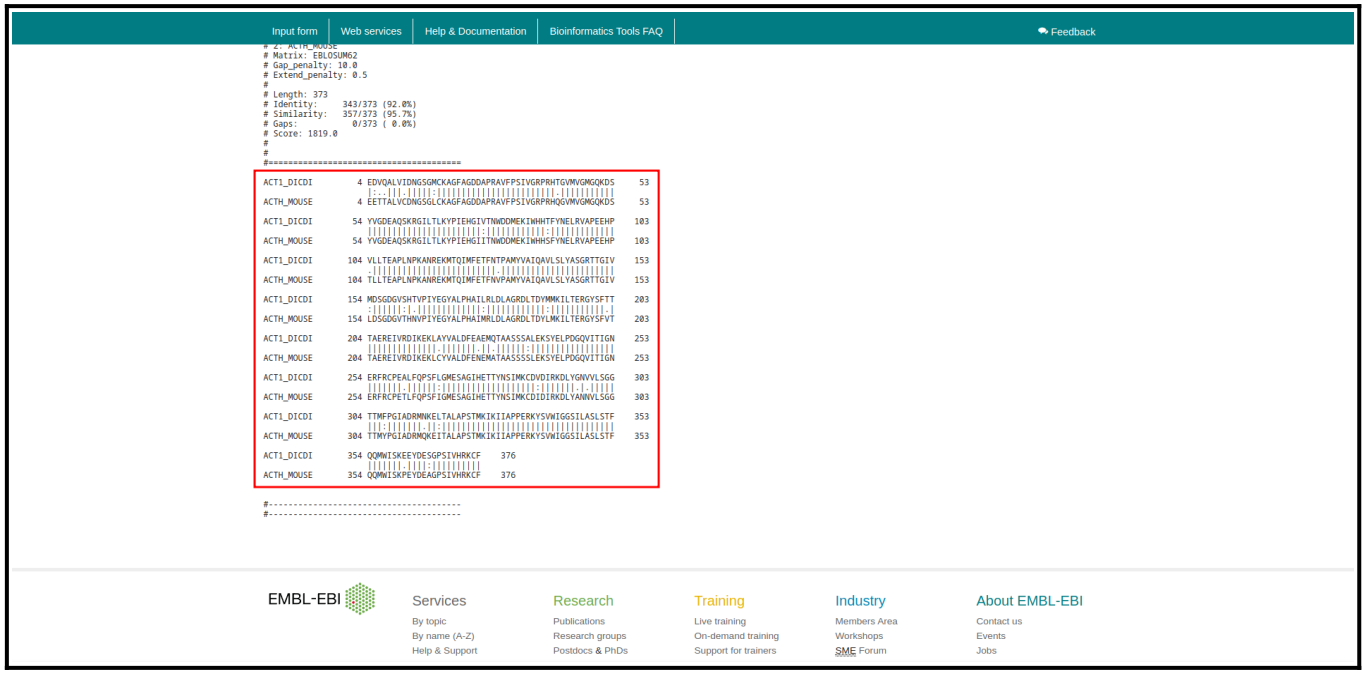

Fig5. Two Sequences Alignment for Actin

Results:

In Smith-Water alignment algorithm for Actin sequences it showing score 1819. It has length 373, Identity 343/373 (92.0%), Similarity 357/373 (95.7%), Gaps 0/373. At matrix EBLOSUM62 with gap penalty10 and extend penalty 0.5.

Conclusion:

The alignment tab shows the alignment of the two sequences, with all the described parameters, used scoring matrices and Gap penalty scored values. The alignment tab has an option fo rhte user to download the entire alignemtn. The alignment between the two sequences is shown in Figure 4, the gaps are represented with '-'. If a match is there between the two nucleotide there is a symbol '|' and the mismatch is represented with a dot '.'

References:

1. Xiong, J. (2006). Essential Bioinformatics (1st ed.). Cambridge University Press.
2. (n.d.). Retrieved from https://www.uniprot.org/uniprot/P63268.fasta
3. (n.d.). Retrieved from https://www.uniprot.org/uniprot/P07830.fasta
4. Dominguez, R., & Holmes, K. C. (2011). Actin structure and function. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3130349/
5. Embl-Ebi. (n.d.). Pairwise Sequence Alignment. Retrieved from https://www.ebi.ac.uk/Tools/psa/
6. Embl-Ebi. (n.d.). EMBOSS Needle. Retrieved from https://www.ebi.ac.uk/Tools/psa/emboss_needle/
7. Embl-Ebi. (n.d.). EMBOSS Stretcher. Retrieved from https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=emboss_stretcher-I20211026-162235-0667-98345526-p1m
8. Embl-Ebi. (n.d.). EMBOSS Stretcher. Retrieved from https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=emboss_stretcher-I20211026-162235-0667-98345526-p1m