

1.EMBOSS needle is predefined with the scoring matrices DNAMfull for nucleotide sequence, BLOSUM65 for protein sequence (Figure 2).

2.The gap open and gap extend penalty can be changed by user defined values. In this example it kept as default values.

3.The user can be notified results through email, if the checkbox has been checked and the mail address is submitted.

The screenshot displays the 'STEP 2 - Set your pairwise alignment options' section of the EMBOSS Needle web interface. It features three dropdown menus: 'MATRIX' set to 'DNAMfull', 'GAP OPEN' set to '10', and 'GAP EXTEND' set to '0.5'. The 'OUTPUT FORMAT' dropdown is set to 'pair'. The 'GAP OPEN' and 'GAP EXTEND' labels and their respective values are circled in red. Below this section is 'STEP 3 - Submit your job', which includes an unchecked checkbox for email notifications and a red 'Submit' button.

Figure 2: Screenshot to set the parameters for the pair wise sequence alignment using Smith –Waterman algorithm

Interpretation:

Once you have clicked on the submit button, the results are displayed within few minutes. The results page comprises of three tabs namely Alignment, Submission details and submit another job. 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 for the user to download the entire alignment file by clicking on the button “View Alignment File”. The submission details tab displays user specified details like the program used, time and date of when the program has launched and the internal commands used for the program execution. Also the user can download the input and output files from this tab .

Aim:

- To study the pairwise sequence similarity search using BLAST algorithm.
- To study the functional and evolutionary relationships between different sequences.

Theory:

BLAST program was designed by Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers, and David J. Lipmann at National Institutes of Health (NIH) and was published in Journal of Molecular Biology in 1990. BLAST (Basic local alignment search tool) is a heuristic search algorithm, it finds the solutions from the all possibilities, which takes input as nucleotide or protein sequence and compare it with existing databases like NCBI, GenBank etc. It finds the local similarity between different sequences and calculates the statistical significance of matches. It can also be used to find functional and evolutionary relationship between different sequences. Search is done by taking the sequence of a certain word size, comparing it with the database sequence and scores are assigned for each comparison. Based on the threshold, a suitable match of that query word is taken and the alignment is extended to both sides. After the alignment is complete, the total score is calculated and alignment is displayed on the blast result page only if the total scores exceed the threshold value.

Sequence, Sequence Alignment and importance:

A biological sequence refers to a sequence of characters which belong to DNA/RNA/protein. Two types of biological sequences are most commonly known namely, Nucleotide Sequence and Protein Sequence. Nucleotide sequence is mainly formed of four different nucleotides namely, adenine (A), guanine (G), cytosine (C) and thymine (T). While protein sequence is formed of 20 different amino acids which are commonly found. The nucleotides arrange themselves in the form of triplet code (triplet code refers to a group of three nucleotides) to code for an amino acid. These sequences are properly indexed in the already existing databases and it is possible to retrieve these sequences from their corresponding databases. The sequences are obtained by the following methods explained below.

DNA sequencing methods:

Sanger Method (dideoxy chain termination method): Here 4 test tubes are taken labelled with A, T, G and C. Into each of the test tubes DNA has to be added in denatured form (single strands). Next a primer is to be added which anneals to one of the strand in template. The 3' end of the primer accommodates the dideoxy nucleotides [ddNTPs] (specific to each tube) as well as the deoxy nucleotides randomly. When the ddNTP's gets attached to the growing chain, the chain terminates due to lack of 3'OH which forms the phospho diester bond with the next nucleotide. Thus small strands of DNA are formed. Electrophoresis is done and the sequence order can be obtained by analysing the bands in the gel based on the molecular weight. The primer or one of the nucleotides can be radioactively or fluorescently labeled also, so that the final product can be detected from the gel easily and the sequence can be inferred.

Maxam-Gilbert (Chemical degradation method): This method requires denature DNA fragment whose 5' end is radioactively labeled. This fragment is then subjected to purification before proceeding for chemical treatment which results in a series of labeled fragments. Electrophoresis technique helps in arranging the fragments based on their molecular weight. To view the fragments, gel is exposed to X-ray film for autoradiography. A series of dark bands will appear, each corresponding to a radio labeled DNA fragment, from which the sequence can be inferred.

Procedure:

This is the common procedure for any BLAST program.

Step 1: Select the BLAST program.

Step 2: Enter a query sequence or upload a file containing sequence.

Step 3: Select the database to search.

Step 4: Select the algorithm and the parameters of the algorithm for the search.

Step 5: Run the BLAST program.

Step 1: Select the BLAST program

User have to specify the type of BLAST programs from the database like BLASTp, BLASTn, BLASTx, tBLASTn, tBLASTx.

Step 2: Enter a query sequence or upload a file containing sequence

Enter a query sequence by pasting the sequence in the query box or uploading a FASTA file which is having the sequence for similarity search. This step is similar for all BLAST programs. The user can give the accession number or gi number or even a raw FASTA sequence. Go to simulator tab to know more about how to retrieve query sequence.

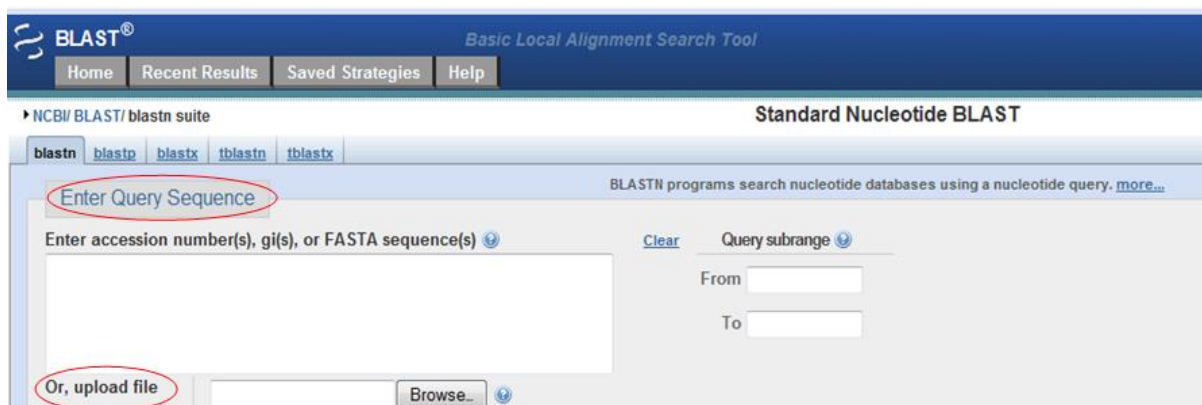


Figure 1: Enter a query sequence or upload a file containing sequence

Step 3: Select database to search

User first has to know what all databases are available and what type of sequences are present in those databases. Sequence similarity search involves searching of similar sequences of the query sequence from the selected databases (Figure 2).

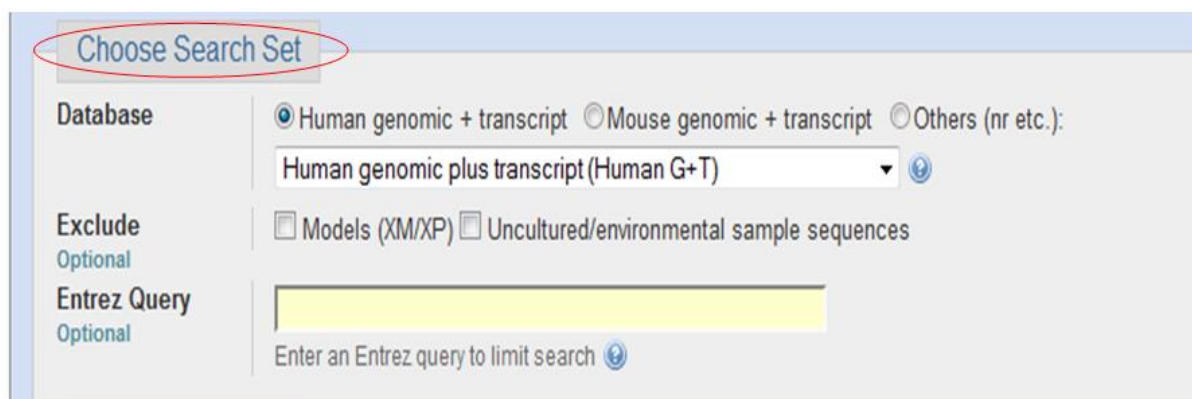


Figure 2: Select database to search

Step 4: Select the algorithm and the parameters of the algorithm for the search

There are different algorithms for some of the BLAST program. User has to specify the algorithm for the BLAST program. Nucleotide BLAST uses algorithms like MegaBLAST which searches for highly similar sequences, discontinuous MegaBLAST which searches for more dissimilar sequences and BLASTn which searches for somewhat similar sequences. Meanwhile for protein BLAST algorithms like BLASTp, searches for similarity between protein query and protein database, PSI-BLAST performs position specific search iteratively, PHI-BLAST searches for a particular pattern (user has to enter the pattern to search in the PHI pattern box provided) that is present in the sequence against the sequences in the database, DELTA-BLAST is Domain Enhanced Lookup Time Accelerated BLAST. It searches multiple sequence and aligns them to find protein homology. The different algorithmic parameters are, Target sequences, Short queries, E-value, Word size, Query range, scoring parameters (Match/Mismatch scores, and Gap penalties) and filters (Filter and Mask) which are required to run BLAST programs. Default values are provided but the user can adjust the values accordingly which is shown in figure 3.

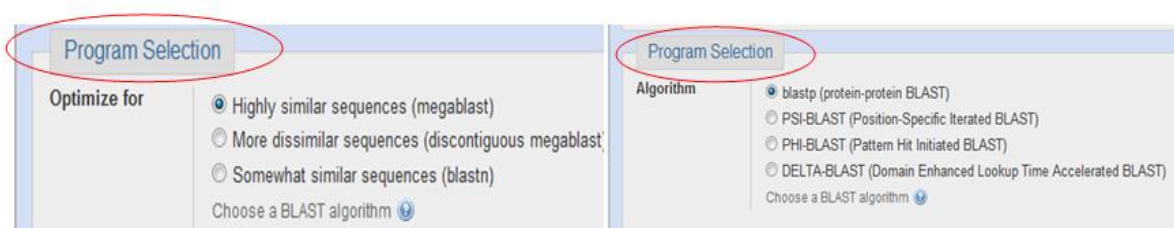


Figure 3: Algorithm and the parameters

Step 5: Run the BLAST program

Submission of the BLAST program can be done by clicking the BLAST button at the end of the page. Screen shot of result can be shown in figure 4.

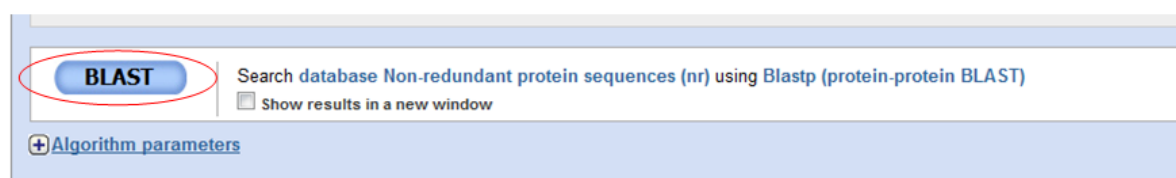


Figure 4: Run the BLAST program

Interpretation:

The query sequence represented as a numbered red bar below the color key. Database hits are shown below the query (red) bar according to the alignment score. Among the aligned sequences, the most related sequences are kept near to the query sequence. User can find more description about these alignments, by dragging the mouse to the each colored bar. The alignment is preceded by the sequence identities, along with the definition line, length of the matched sequence, followed by the score and E-value. The line also contains the information about the identical residues in alignment (identities), number of positivity's, number of gaps used in the alignment. Finally it shows the actual alignment, along with the query sequence on the top and database sequence below the query. The number on either sides of the alignment indicates the position of amino acids/nucleotides in sequence.

Experiment 4: Pairwise sequence alignment using FASTA

Aim: FASTA can carry out a dynamic sequence similarity search between the Protein and Nucleotide sequences against the databases.

Theory: FASTA is a pairwise sequence alignment tool which takes input as nucleotide or protein sequences and compares it with existing databases. It is a text-based format and can be read and written with the help of text editor or word processor. Fasta file description starts with '>' symbol and followed by the gi and accession number and then the description, all in a single line. Next line starts with the sequence and in each row there would be 60 nucleotides/amino acids only. For DNA and proteins it is represented in one letter IUPAC nucleotide codes and amino acid codes. It finds the local similarity between the sequences and calculates the statistical significance of matches. It can be also used to find the functional and evolutionary relationship between the sequences.

FASTA program uses the word hits to identify potential matches before attempting the more time consuming optimised search. The speed and sensitivity is controlled by the parameter called ktup, which specifies the size of the word. Increasing the ktup decreases the number of background hits. Initially it checks for segment's containing several nearby hits. This program is much more sensitive than BLAST programs, which is reflected by the length of time required to produce results. FASTA produces local alignment scores for the comparison of the query sequence to every sequence in the database. This approach avoids the artificiality of a random sequence model by real sequences, with their natural correlations. The sequences are obtained by the following methods.

Procedure:

There are four steps require to run FASTA program.

Step 1: Specify the tool input (sequence and database).

Step 2: Entering of input sequence.

Step 3: Set up the parameters.

Step 4: Submit the query for processing.

Use this tool

STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected ✕ Clear Selection

- ☒ UniProt Knowledgebase
- ☐ UniProtKB/Swiss-Prot
- ☐ UniProtKB/Swiss-Prot isoforms
- ☐ UniProtKB/TrEMBL
- ☐ UniProtKB Taxonomic Subsets

OTHER TYPES

General

- Nucleotide Databases

Specialised

- Proteomes Databases
- Genomes Databases
- WGS Databases

STEP 2 - Enter your input sequence

Enter or paste a PROTEIN sequence in any supported format:

or Upload a file: Browse...

STEP 3 - Set your parameters

PROGRAM

FASTA

MATRIX	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE
BLOSUM50	-10	-2	2	10	0 (default)
DNA STRAND	HISTOGRAM	FILTER	STATISTICAL ESTIMATES		
N/A	no	none	Regress		
SCORES	ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE	MULTI HSPs	
50	50	START-END	START-END	no	
SCORE FORMAT					
Default					

STEP 4 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

Submit

Figure 1: Fasta home page

Step 1: Specify the tool input

Select the database to search :Databases are required to run the sequence similarity search. Multiple databases can be used at the same time. The different databases are

Uniprot Knowledge base

Uniprot KB/swiss-prot

Uniprot KB/ Swissprot isoforms

Uniprot KB /Trembl

UniProtKB Taxonomic Subsets

UniProt Clusters

Patents

Structure



Figure 2 : Selecting the database

Step 2 Entering of input sequence

The query sequence can be entered directly in GCG, FASTA, EMBL, GenBank, PIR, NBRF, PHYLIP or UniProtKB/Swiss-Prot formats.

Sequence file upload

A file containing the valid sequence in any format mentioned above can be used as a query for sequence similarity search. Sequence type indicates the type of sequence (PROTEIN / DNA / RNA) for similarity search. Go to simulator tab to know more about how to retrieve the query sequence.

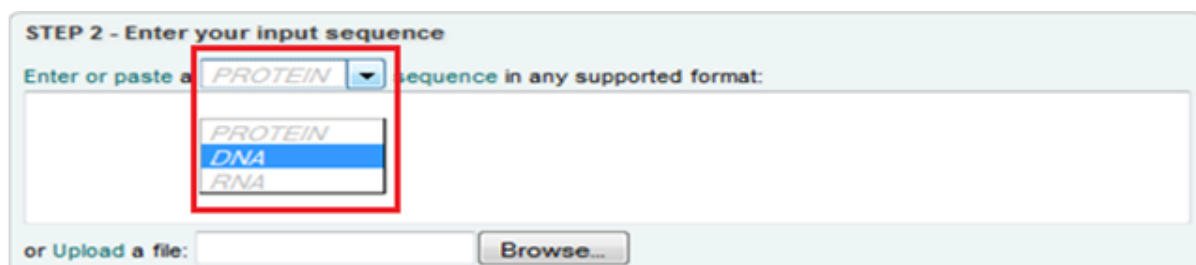


Figure 3 : Entering of input sequence

Step 3: Setting up parameters

User has to specify the type of program and the matrix for scoring. FASTA, FASTX, FASTY, SSEARCH, GGSEARCH and GLSEARCH are the different programs used. Substitution matrix are used for scoring alignments. The matrices are BLOSUM50, BLOSUM62, BLASTP62, PAM120, PAM250, MDM10, MDM20, and MDM40. BLOSUM50 is set as a default substitution matrix. Parameters include.

GAP open and GAP extended penalty: Common and regular cause for GAP is mutation, if gap penalty is low we can get high scoring sequence similarity search. Also gaps will increase uncertainty in alignment.

Ktup: It is a value given as the word size for comparison.

Expectation value (E-value): It decreases exponentially with the score that is assigned to an alignment between two sequences.

Strands, Histograms, Filter: It filters the low complex regions in sequence similarity search. Histogram will give graphical representation of scores.

Statistical estimates, Scores, alignments, sequence range and database range: specify the range of the query for search in database.

HSPs, Score format, Transition table score format: are the different score formats. Transition table gives the genetic codes used in translation.

STEP 3 - Set your parameters

PROGRAM
FASTA ▼

MATRIX	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE
BLOSUM50 ▼	-10 ▼	-2 ▼	2 ▼	10 ▼	0 (default) ▼

DNA STRAND	HISTOGRAM	FILTER	STATISTICAL ESTIMATES
N/A ▼	no ▼	none ▼	Regress ▼

SCORES	ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE	MULTI HSPs
50 ▼	50 ▼	START-END	START-END	no ▼

SCORE FORMAT **TRANSLATION TABLE**

Default ▼ N/A ▼

Figure 4 : Setting up parameters.

Step 4: Submission

The result page can be seen in another window by clicking submit. This is an interactive process, when the process is complete the result will be displayed in the browser. Result can be sent to a valid email address which has to be specified in the text box.

STEP 4 - Submit your job
☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

Submit

STEP 4 - Submit your job
☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

Submit

STEP 3 - Set your parameters

PROGRAM
FASTA

MATRIX	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE
BLOSUM50	-10	-2	2	10	0 (default)

DNA STRAND	HISTOGRAM	FILTER	STATISTICAL ESTIMATES
N/A	no	none	Regress

SCORES	ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE	MULTIHSPs
50	50	START-END	START-END	no

SCORE FORMAT
TRANSLATION TABLE
Default
N/A

Figure 5 :Submission

Interpretation:

The tool output will be giving you complete statistical details of the sequence similarity search. Visual output that is the FASTA visual output gives the result of the sequence match and subject match with their E-values in a colour full schema.

Experiment 5: Aligning Multiple Sequences with CLUSTAL W

Aim: To align three or more sequences to find out structural and functional relationship between these sequences.

Theory: Sequence is a collection of nucleotides or amino acid residues which are connected with each other. Speaking biologically, a typical DNA/RNA sequence consist of nucleotides while a protein sequence consist of amino acids.

Sequencing is the process to determine the nucleotide or amino acid sequence of a DNA fragment or a protein. There are different experimental methods for sequencing, and the obtained sequence is submitted to different databases like NCBI, Genbank etc.

Procedure:

Steps to perform multiple sequence alignment:

To download the data , and to get acces to the tools, go to simulator tab.

Get access to the CLUSTALW tool

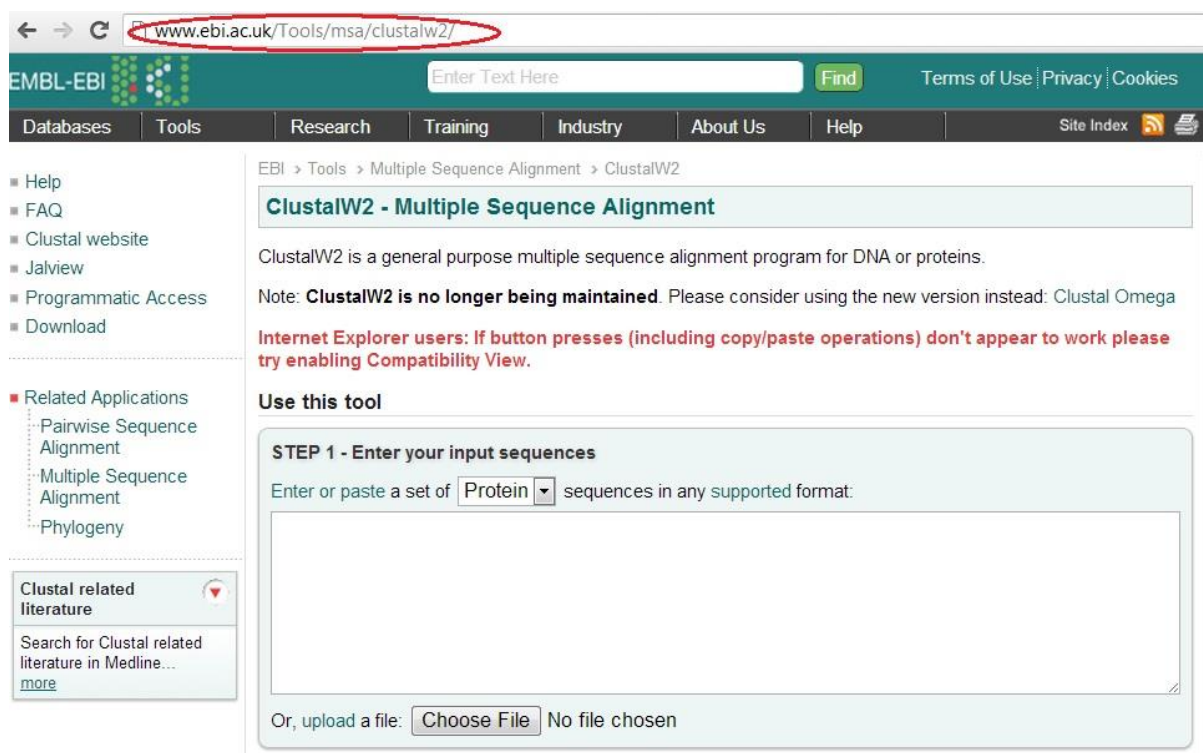


Figure 1: Screenshot of the CLUSTALW tool

In the dialog box given, paste your set of sequences, the sequences should be pasted with the '>' symbol followed by name of the sequence (as similar as FASTA format) followed by return (enter key) and then the sequence (Figure 2).

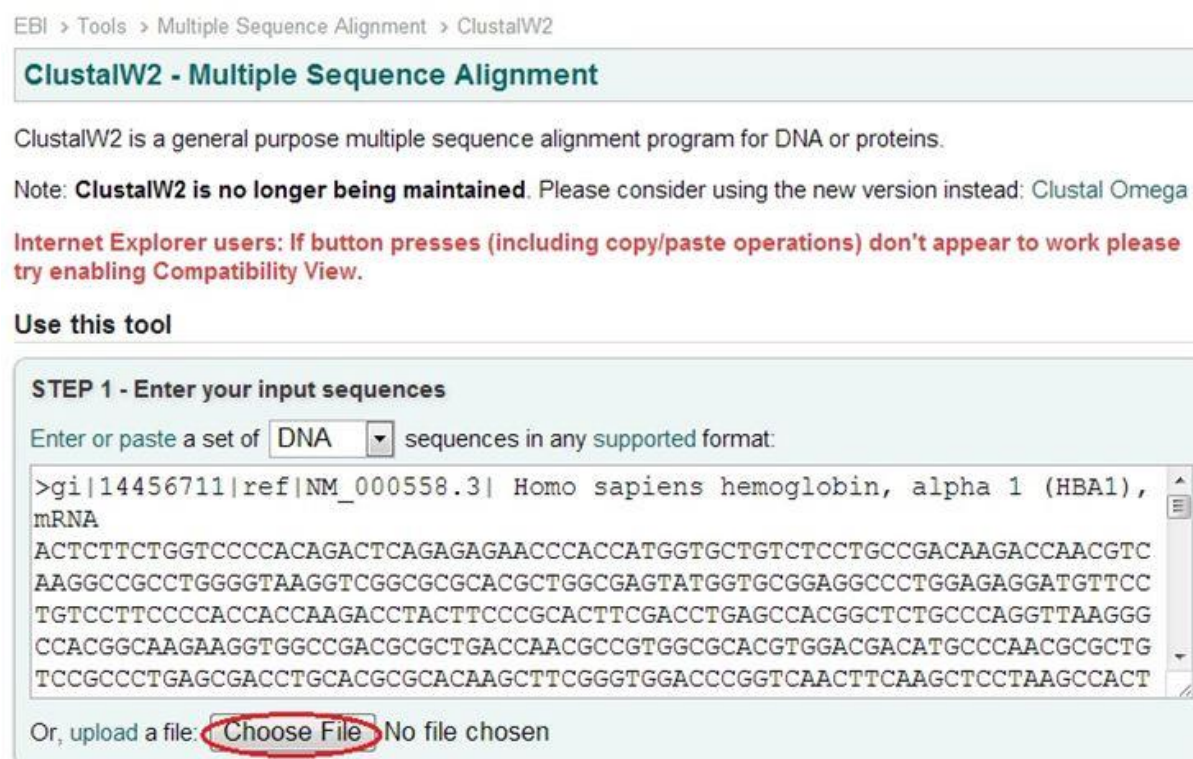


Figure 2: Screenshot to paste the sequence for alignment

The sequences can also be submitted through file by clicking on the option “choose file” such that all the sequences should be in similar format.

The other two steps the user can select on his/her own to set the parameters for pair wise alignment options and multiple sequence alignment options, to select the scoring matrices and scoring values. In most of the cases the parameters are set default (Figure 3).

Results can be notified by email when the user checks the button email notification (Figure 3).

STEP 2 - Set your Pairwise Alignment Options
Alignment Type: ☒ Slow ☐ Fast
Slow Pairwise Alignment Options
DNA Weight Matrix: ClustalW GAP OPEN: 10 GAP EXTENSION: 0.1

STEP 3 - Set your Multiple Sequence Alignment Options
DNA Weight Matrix: ClustalW GAP OPEN: 10 GAP EXTENSION: 0.20 GAP DISTANCES: 5 NO END GAPS: no
ITERATION: none NUMITER: 1 CLUSTERING: NJ
OUTPUT Options
FORMAT: Aln w/numbers ORDER: aligned

STEP 4 - Submit your job
☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)
Submit

Figure 3: Screenshot of the Parameters to be submitted for the alignment

After the submission of the job the results can be downloaded into a file by clicking on the option Download alignment file (Figure 4).

ClustalW2 Results
Alignments Result Summary Guide Tree Submission Details Submit Another Job

Alignment
Download Alignment File

CLUSTAL 2.1 multiple sequence alignment

```

gi|14456711|ref|NM_000558.3|      -----
gi|224809341|ref|NG_011508.1|    --CCGATTTACACATACAATA-----GGGATGAATCTCA  32
gi|196115071|ref|NG_008356.1|    GTCTGACATGTTTTTAAATATTTTGTTCCTAGTGTGTGATGCAATTCT  50

gi|14456711|ref|NM_000558.3|      -----
gi|224809341|ref|NG_011508.1|    AAATAATCATGCTGAGTAAAAGAAACCAGGAGAAAAAAG--TAGATGCC  80
gi|196115071|ref|NG_008356.1|    TTGCAATCTT----ATTAAAGTCTACTATGCCCCACAAATGCTTAGAAGGC  96

gi|14456711|ref|NM_000558.3|      -----
gi|224809341|ref|NG_011508.1|    AT--ATTATCTCACTTAC-ATAAAATCTGGAAAAAT---ACAACTAATC  124
gi|196115071|ref|NG_008356.1|    ATGGATT-TCGCACATATTATCAGGTGCCACTGGATCTAACAGTGTACTT  145

gi|14456711|ref|NM_000558.3|      -----
gi|224809341|ref|NG_011508.1|    TAGTGTGACAGAAAGGAGATCAGCGGCTTCCAGGAGATAAG-AATGGAGA  173
gi|196115071|ref|NG_008356.1|    GACTGTG-TAGATAAAAGGTCA-TGACATCCAGGAGTTTAGTAAAGGAGA  193

```

Figure 4: Screenshot to download the alignment file

The result summary tab gives the links to different outputs summary and link to each output (Figure 5).