Let’s run multiple sequence alignments using the clustalOmega algorithm, which is also known as clustalO.

To Enter the program, use the following link:

http://www.ebi.ac.uk/Tools/msa/clustalo/

We’ll compare the c-fos protein as found in different organisms. The program can accept sequences in different formats as input. We will use the FASTA format which is the most common one. In this format the first line is a description line that should start with a greater than symbol “>” followed by any description, while the sequence itself appears from the second line onwards.

Enter all sequences to the input text box.

```
>FOS_CHICKEN Proto-oncogene protein c-fos (Cellular oncogene
fos).
MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANF
VPTVTAISTSPDLQWLVQPTLISSVAPSQNRGHPYGVPAPAPPAAYSRPAVLKAPGGRGQ
SIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRNRRRELTDTLQAETDQLEEEKSALQA
EIANLLKEKEKLEFILAAHRPACKMPEELRFSEELAAATALDLGAPSPAAAEEAFALPLM
TEAPPAVPPKEPSGSGLELKAEPFDELLFSAGPREASRSVPDMDLPGASSFYASDWEPLG
AGSGGELEPLCTPVVTCTPCPSTYTSTFVFTYPEADAFPSCAAAHRKGSSSNEPSSDSLS
SPTLLAL
>FOS_CARP Proto-oncogene protein c-fos (Cellular oncogene fos).
MMFTSLNADCDASSRCSTASAAAESVACYPLNQTQKFTELSVSSASFVPTVTAISSCPDL
QWMVQPMVSSVAPSNGGARSYNPNPYPKMRVTGTKSPNSNKRARAEQLSPEEEEKKRVRR
ERNKMAAAKCRNRRRELTDTLQAETDELEDEKSALQNDIANLLKEKERLEFILAAHKPIC
KIPSSSVSPIPAASVPEIHSITTSVVSTANAPVTTSSSSSLFSSTASTDSFGSTVEISDL
EPTLEESLELLAKAELETARSVPDVDLSSSLYARDWESLYTPANNDLEPLCTPVVTRTPA
CTTYTSSFTFTYPENDVFPSCGPVHRRGSSSNDQSSDSLNSPTLLTL
>FOS_HUMAN Proto-oncogene protein c-fos (Cellular oncogene fos)
(G0/G1 switch regulatory protein 7).
MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNAQDFCTDLAVSSANF
IPTVTAISTSPDLQWLVQPALVSSVAPSQTRAPHPFGVPAPSAGAYSRAGVVKTMTGGRA
QSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRNRRRELTDTLQAETDQLEDEKSALQ
TEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEMSVASLDLTGGLPEVATPESEEAFT
LPLLNDPEPKPSVEPVKSISSMELKTEPFDDFLFPASSRPSGSETARSVPDMDLSGSFYA
ADWEPLHSGSLGMGPMATELEPLCTPVVTCTPSCTAYTSSFVFTYPEADSFPSCAAAHRK
GSSSNEPSSDSLSSPTLLAL
>FOS_HAMSTER Proto-oncogene protein c-fos (Cellular oncogene
fos).
MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNAQDFCTDLSVSSANF
IPTVTAISTSPDLQWLVQPTLVSSVAPSQTRAPHPYGVPTPSTGAYSRAGMVKTVSGGRA
QSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRNRRRELTDTLQAETDQLEDEKSALQ
TEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEMFVASLDLTGGLPEATTPESEEAFS
LPLLNDPEPKPSLEPVKSISNVELKAEPFDDFLFPASSRPSGSETTARSVPDMDLSGSFY
AADWEPLHSSSLGMGPMVTELEPLCTPVVTCTPSCTTYTSSFVFTYPEADSFPSCAAAHR
KGSSSNEPSSDSLSSPTLLAL
>FOS_MOUSE Proto-oncogene protein c-fos (Cellular oncogene fos).
MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF
IPTVTAISTSPDLQWLVQPTLVSSVAPSQTRAPHPYGLPTQSAGAYARAGMVKTVSGGRA
QSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRNRRRELTDTLQAETDQLEDEKSALQ
TEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEMSVASLDLTGGLPEASTPESEEAFT
LPLLNDPEPKPSLEPVKSISNVELKAEPFDDFLFPASSRPSGSETSRSVPDVDLSGSFYA
ADWEPLHSNSLGMGPMVTELEPLCTPVVTCTPGCTTYTSSFVFTYPEADSFPSCAAAHRK
GSSSNEPSSDSLSSPTLLAL
>FOS_PIG Proto-oncogene protein c-fos (Cellular oncogene fos).
MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNAQDFCTDLAVSSVNF
IPTVTAISISPDLQWLVQPTLVSSVAPSQTRAPHPYGVPTPSAGAYSRAGAVKTMPGGRA
QSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRNRRRELTDTLQAETDQLEDEKSALQ
TEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEMSVASLDLSGGLPEAATPESEEAFT
LPLLNDPEPKPSVEPVKKVSSMELKAEPFDDFLFPASSRPGGSETARSVPDMDLSGSFYA
ADWEPLHGGSLGMGPMATELEPLCTPVVTCTPSCTAYTSSFVFTYPEADSFPSCAAAHRK
GSSSNEPSSDSLSSPTLLAL
>FOS_PUFFERFISH Proto-oncogene protein c-fos (Cellular oncogene
fos).
MMFTSFNAECDSSSRCSASPVGDNLYYPSPAGSYSSMGSPQSQDFTDLTASSASFIPTVT
AISTSPDLQWMVQPLISSVAPSHRAHPYSPSPSYKRTVMRSAASKAHGKRSRVEQTTPEE
EEKKRIRRERNKQAAAKCRNRRRELTDTLQAETDQLEDEKSSLQNDIANLLKEKERLEFI
LAAHQPICKIPSQMDTDFSVVSMSPVHACLSTTVSTQLQTSIPEATTVTSSHSTFTSTSN
SIFSGSSDSLLSTATVSNSVVKMTDLDSSVLEESLDLLAKTEAETARSVPDVNLSNSLFA
AQDWEPLHATISSSDFEPLCTPVVTCTPACTTLTSSFVFTFPEAETFPTCGVAHRRRSNS
```

Please note that it is important to give informative names at the beginning of the description line, otherwise we could not differentiate sequences at the output. The simplest way to do it is by adding the gene name or the organism name after the "greater than" symbol > in each description line.

Choose the "protein" option under step1.

Choose the "clustalw with characters counts" option under step2 for presenting Clustal alignment format with residue numbering.

Click on Submit.

At the output page you'll find several output files created by the program. They'll be organized in different tabs.

All running files and the percent identity matrix appear under the "Result Summary" tab.

Click on the "Percent Identity Matrix" link. ClustalO first calculates the percent of identity for each pair of two sequences and presents it in this matrix. The titles of the columns are the same as in the rows but just transposed. The most conserved pair of sequences in this matrix are the mouse and the hamster ones with about 96% identity. The pig and the human sequences have also a high identity value of 96%.

The multiple sequence alignment appears under the "Alignments" tab.

Since most sequences are longer than the 60 characters limit that can be presented in a line, the alignment is spread over to several blocks of 60 characters each. The sequence of each organism will appear at the same line in each block. For example, the chicken sequence appears in the first block as the first sequence in the alignment. Its last amino acid in this block is amino acid 60, whereas amino acid 61 appears in the next block on the left side of the first line. Please note that the last amino acid in this line is 119 and not 120 as expected because there is a gap in the middle of it.

The number of the last amino acid in each organism appears at the right side of the last block, representing the length of each original sequence.

We evaluate the MSA by looking at the conservation of each column in the alignment. There are consensus signs under each column:

An Asterisk "*" is used for columns in which all amino acids are identical

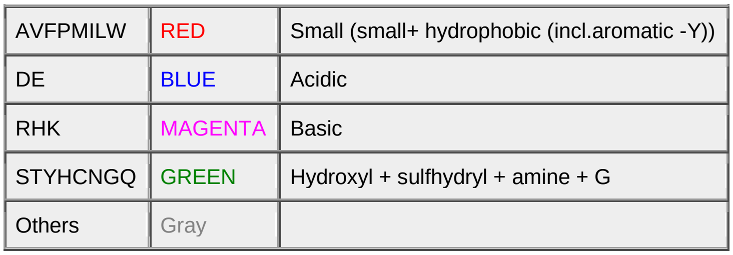
A colon/ two dots, ":" is used for highly conserved columns, showing only very similar amino acids. The similarity between two amino acids is defined by a similarity matrix like blosum62.

A period/one dot, "." is used for Partially conserved columns, that can include also less similar amino acids.

No symbol will appear under a column if it is not conserved at all.

An area enriched with Asterisks and colons represents a high conserved common motif, whereas, areas with no symbols or areas enriched with gaps represent non-conserved regions.

Click on "Show colors" to present alignment with the help of distinct colors for each group of amino acids. The meaning of the standard color scheme is as follows:



AVFPMILW	RED	Small (small+ hydrophobic (incl.aromatic -Y))
DE	BLUE	Acidic
RHK	MAGENTA	Basic
STYHCNGQ	GREEN	Hydroxyl + sulfhydryl + amine + G
Others	Gray	

For example, acidic amino acids will be colored in blue, whereas basic ones will be colored in magenta. In order to remove the colors, click on the “Hide colors”.

Now we’ll learn how to create a Logo from the MSA output. We will use WebLogo, a web based application designed to generate sequence logos. Sequence logos are graphical representations of the bottom line of an amino acid or nucleic acid multiple sequence alignment. Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position. In general, a sequence logo provides a richer and more concise description of, for example, a binding site, than would a consensus sequence.

There are three MSA input formats that can be used in Weblogo; FASTA, CLUSTALW (the default output format of ClustalO), and flat. All sequences must have the same length.

No digits are permitted at the input; therefore we need to run ClustalO again , this time changing the output format parameter (under step2) to ClustalW without the numbering output we requested before .

Enter to the WEBLOGO page using the following link:

http://weblogo.berkeley.edu/logo.cgi

Paste the Clustalo MSA output. Don’t forget to include also the consensus signs under all columns.

V check the following parameter in order to get a clear output:

Multiline Logo (Symbols per Line): 32

Enter 124-207 at the “Logo Range” parameter.

Click on Create Logo.

Now you’ve generated a nice Logo which can be presented as a conserved motif figure in an article.

Do note that a position in the logo in which no letter is displayed means that there is no conservation in that position. This could happen due to gaps or variety of amino acid at that position.

You should note that the clustal family of algorithms has many installations on many sites, and user interface may not be identical in all of them. Furthermore, the interface may change and get updated from time to time. Hopefully today's hands-on experience will help you navigate this very useful program in the future as well.