

Multiple Sequence Alignments

Recap:

A multiple sequence alignment is one of the most important investigative tools you can generate in Bioinformatics. A multiple sequence alignment is useful for predicting protein structure, function and is intrinsic to phylogenetic analysis.

Aim of Practical:

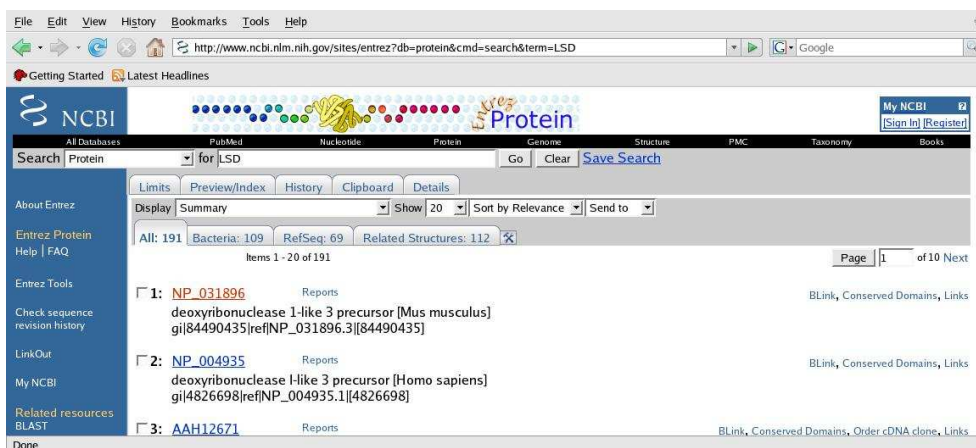
The aim of this practical is to get you familiar with keyword searching for sequences of interest, doing homology searches using BLAST and aligning sequences using two different alignment applications.

Part 1: NCBI searching

Go to the NCBI website: <http://www.ncbi.nlm.nih.gov/>

(Later, if you wish, take some time to click on the various links on this website.

It is a hub of information.) Choose the Protein database from the 'All Databases' drop-down window. Type the keyword 'LSD' into the search window and hit 'Go'. The resulting page should look something like this

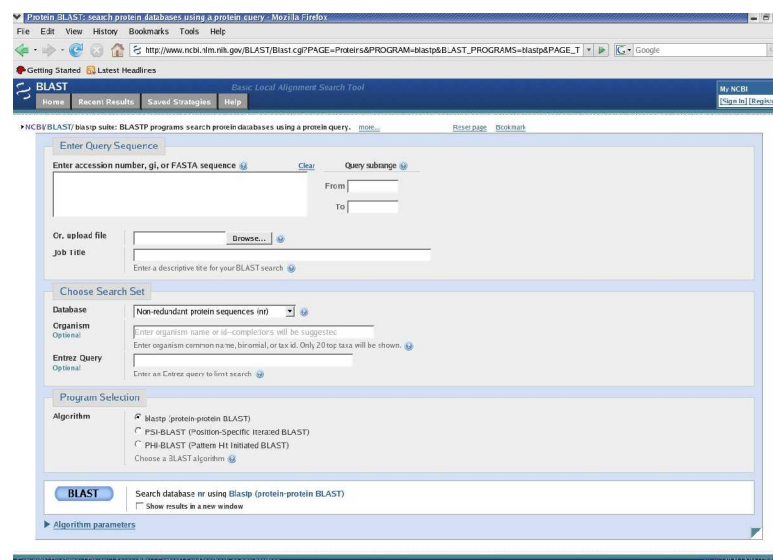


Look through the results page and notice how many different functions / species resulted with this simple search word. Click on any of the links (which are unique identification keys for that sequence) if you wish, and you will find a list of information with respect to that sequence, within that species (such as authors, journal publication, chromosomal location etc), as well as the sequence itself (at the bottom of the page).

Wildcards (*) can also be used when performing a keyword search, to enable searches such as immuno*, haemo* etc. This will result in all results beginning with these letters. Boolean operators, AND, NOT and OR are also available for use. Boolean search operators can be used to refine queries; e.g. "**LSD AND mus AND musculus**" will only return results containing all of these words. Try it.....

Part 2: BLAST

BLAST allows for both nucleotide and amino acid searches. As the method for both is virtually identical, we shall just use the protein BLAST tool today. From the initial NCBI page (you may need to click on the 'NCBI' to get back there), click on the link 'BLAST'. You should get a page looking like this:



Click on the link 'protein blast'. Open the file "**REIS_TODPA**" (an opsins from Japanese Flying Squid) and copy the sequence into the query sequence window. Alternatively, use the '**browse**' button to load the sequence directly from the file.

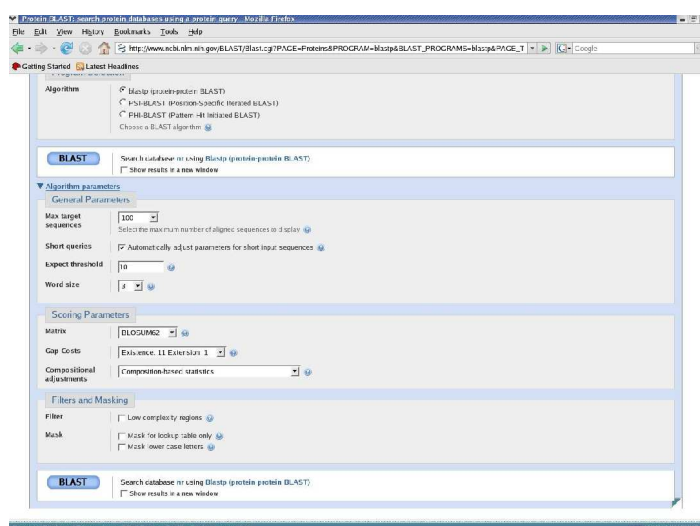
Change the Database to '**Non-redundant protein sequences (nr)**'. This is the largest resource. The alternate databases are subsets of this, which are of use if you have a particular query in mind, for example, if you wish to find homologues with a known structure only, you could use the PDB database.

Do not change any of the other parameters, and hit the BLAST button at the end of the page. The BLAST search will take a short while. Several factors affect the search

time, such as sequence length, sequence complexity and how many other people are using the server at the same time as you!

Now follow the same procedure using the sequence file "*low_complexity.seq*". This sequence contains low complexity, which is defined as a region of a protein sequence enriched with a single amino acid, in this case Glutamine (Q). Do a BLAST search twice, once as previously described and the second time (open another BLAST window if necessary), masking the low complexity regions.

To mask low complexity regions, in the original BLAST window, click the '**Algorithm parameter**' button. This new section allows you to alter the parameters of your search.



Click on the button to mask low complexity regions. This means that those regions will be ignored. Re-run the BLAST search again. A quick glance at the results will show differences from the un-masked sequence. Remember, BLAST is a *local alignment* tool; therefore it will only align *sections* of the sequence. Hence, masking part of a sequence can possibly alter the results. There is no right or wrong way set up your sequence to use with BLAST, just remember that parameter changes can change your results.

Scrolling down past the list of hits, the query sequence aligned to the hit sequences should be found. At this point (though we shall not today), all or some of the resulting sequence hits can be saved to a text file and used as the input file to a multiple sequence alignment program.

Part 3: Multiple Sequence Alignments

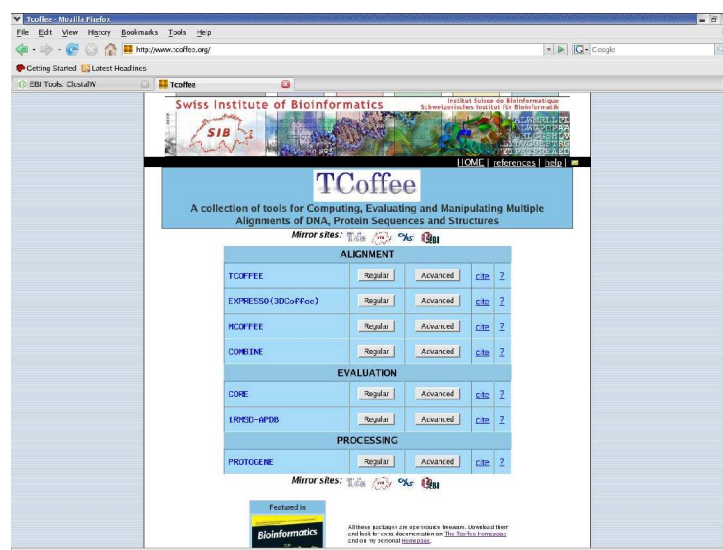
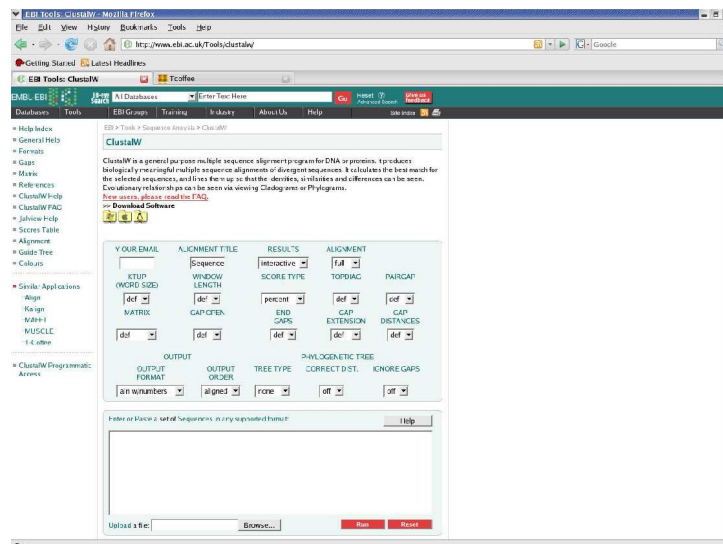
Like a BLAST search, the application used to create an alignment can also affect the result. The two alignment applications demonstrated in the lecture shall be used here, ClustalW and T-COFFEE.

Open an internet window for each of these websites:

<http://www.ebi.ac.uk/Tools/clustalw/>

<http://www.tcoffee.org/>

The two alignments windows should look like this:



There are 2 sequence files to be aligned. The first file "*thio_aln.txt*" contains 16 thioredoxin sequences that show a high level of conservation. The resulting alignment should show this high level of similarity between the underlying sequences.

The second file, "*hard_align.seq*" contains four sequences from prokaryotic species. Previous work (not included here) has shown that the folded structure of these proteins is very similar; however their sequences are very different. This example is to highlight that multiple sequence alignment can get it wrong!

Sequence File 1:

ClustalW

Either open and copy the data from the file "*thio_aln.txt*" or load it into the clustalw window. Make sure that the 'OUTPUT ORDER' is set to 'INPUT'. This will make it easier to compare output of the two methods.

T-COFFEE

Click on the '**Advance**' button, and copy the same sequence file. Make sure the output order is also set to '**INPUT**'. Again, this is just so allow for easy comparison between the two alignment methods.

Once the alignments have completed (after pressing the '**submit**' / '**run**' button), compare the two alignments (click on the **clustalw_aln** button in t-coffee). What do you see? The sequences are very similar, so the alignments are almost identical. Notice there are a few slight differences.

Sequence File 2:

Load in the sequence file "*hard_align.seq*" as described previously and re-run both alignment applications on this set of sequences. With this alignment, you should notice how different the sequences are from each other, how many gaps have to be

places try to align them. Also, there is a more significant difference between the alignments produced by the two methods, highlighting how much the method used can affect the alignment produced.

Conclusion:

From this practical you should have a beginning knowledge of searching the GenBank database located at NCBI. You should be able to now search GenBank, understand the resulting hits, and be able to produce multiple alignments using online tools.

Many other biological databases exist such as, 3 of the main ones are

Uniprot: <http://www.ebi.uniprot.org/index.shtml>

Swissprot: <http://expasy.org/sprot/>

Ensembl : <http://www.ensembl.org/index.html>

Many alignment applications, such as CLUSTALW, are available for downloading to your computer, to remove the need to use the web every time you wish align something.