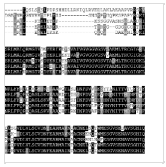


May 20, 2020

Sequence alignment for Biochemistry I

Belle Houston[1], Chris Berndsen[1]

[1]James Madison University

1 *Works for me* This protocol may be deleted by the owner

Chris Berndsen
James Madison University

ABSTRACT

Comparing DNA or protein sequences can provide insight into the structure or function of newly discovered or characterized open reading frames or proteins. The amino acid sequence is often **conserved** between species or between proteins with similar structural/functional properties. A sequence that is conserved means that the sequence is identical across several species. A sequence can have aspects that are similar meaning that while the amino acids are not identical, they have similar properties. Examples of similar sequences are an D --> E change or L --> I change.

The video below provides a brief overview of the process:

MATERIALS TEXT

Microsoft Word or a similar text editing software
Internet connection
A target sequence in FASTA format or a PDB structure or a Uniprot ID

Do you have a sequence already?

1 If you are getting the sequence from Uniprot, go to Step 2

If you are extracting the sequence from a PDB file, go to Step 3

If you have a sequence already provided skip to Step 4.

Obtain Sequence from Uniprot

2 Obtain sequence from **Uniprot** OR extract sequence from structure in **YASARA** (skip to step 3) 1m

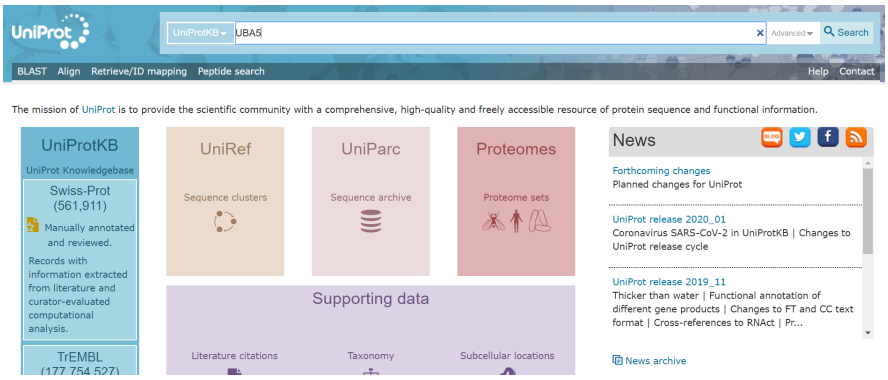
Some things to note about the sequence you use:

- If you obtain sequences from Uniprot, these are full sequences of proteins where the structure may not be entirely known for all amino acids.
- If you extract the sequence from YASARA, the sequence will most likely have less amino acids than the Uniprot sequence as it is only providing the seqeunce of amino acids in the known structure.
- No matter which method you use, be sure to know the species that the sequence originates from.
- For Video Summary see abstract

2.1 1m

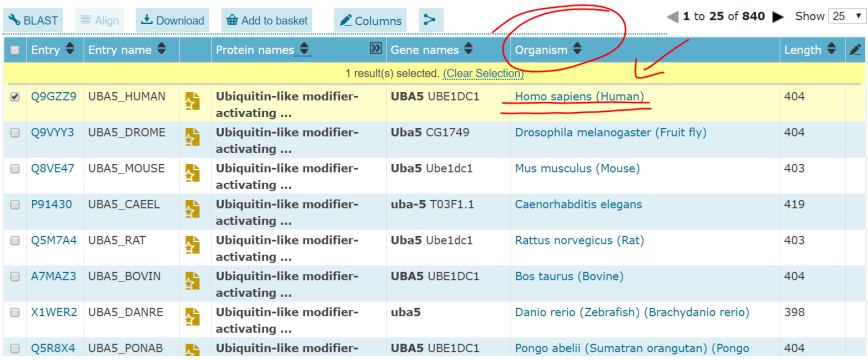
For Uniprot click here

Click on the link above and locate your protein on Uniprot by using the search bar. In this procedure, we will be using the UBA5 protein found in humans as an example protein.



This what you should see for the Uniprot homepage with the name of desired protein typed in the search bar.

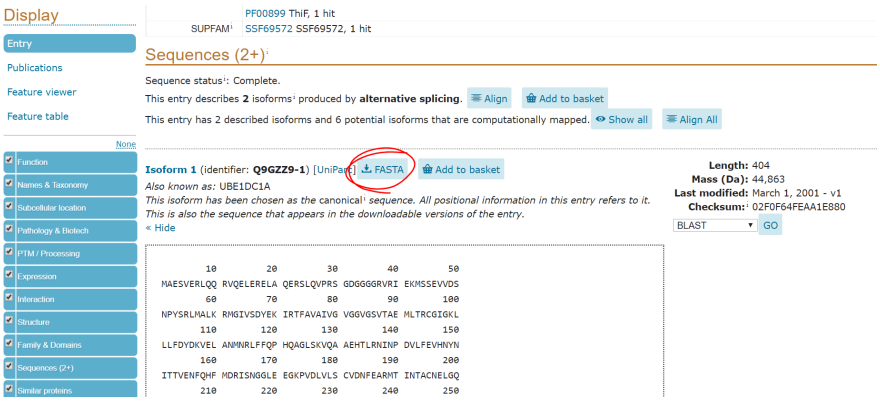
2.2 Your search results will pull up proteins from multiple oranisms. Be sure to chose the protein from the correct organism. In this case, we are interested in the UBA5 in humans. If your protein is found in a plant such as *Arabisopsis*, be sure it says that in the organism column.



Entry	Entry name	Protein names	Gene names	Organism	Length
Q9GZZ9	UBA5_HUMAN	**Ubiquitin-like modifier-activating ...**	**UBA5** UBE1DC1	Homo sapiens (Human)	404
Q9VYY3	UBA5_DROME	**Ubiquitin-like modifier-activating ...**	**Uba5** CG1749	Drosophila melanogaster (Fruit fly)	404
Q8VE47	UBA5_MOUSE	**Ubiquitin-like modifier-activating ...**	**Uba5** Ube1dc1	Mus musculus (Mouse)	403
P91430	UBA5_CAEEL	**Ubiquitin-like modifier-activating ...**	**uba-5** T03F1.1	Caenorhabditis elegans	419
Q5M7A4	UBA5_RAT	**Ubiquitin-like modifier-activating ...**	**Uba5** Ube1dc1	Rattus norvegicus (Rat)	403
A7MAZ3	UBA5_BOVIN	**Ubiquitin-like modifier-activating ...**	**UBA5** UBE1DC1	Bos taurus (Bovine)	404
X1WER2	UBA5_DANRE	**Ubiquitin-like modifier-activating ...**	**uba5**	Danio rerio (Zebrafish) (Brachydanio rerio)	398
Q5R8X4	UBA5_PONAB	**Ubiquitin-like modifier-**	**UBA5** UBE1DC1	Pongo abelii (Sumatran orangutan) (Pongo	404

Selected is UBA5 in humans. Pay attention to the organism column.

2.3 You should be able to locate avaliable sequences of your protein by looking at the options on the left and **clicking the 'sequences tab'**. Or, scroll down until you find the seqeunces. The available sequences will be shown. Sometimes they will provide different isoforms of the protein. Be sure to research which isoform you are interested in if there are different isoforms. **Press the 'FASTA'** download button as shown above the seqeunce.



FASTA should be listed above the displayed sequence.

2.4 Once you press the FASTA download button, a page like the one shown below should come up. **Copy the ENTIRE text starting at the >**. This text is to be pasted into the NCBI pBLAST decribed in the next steps.

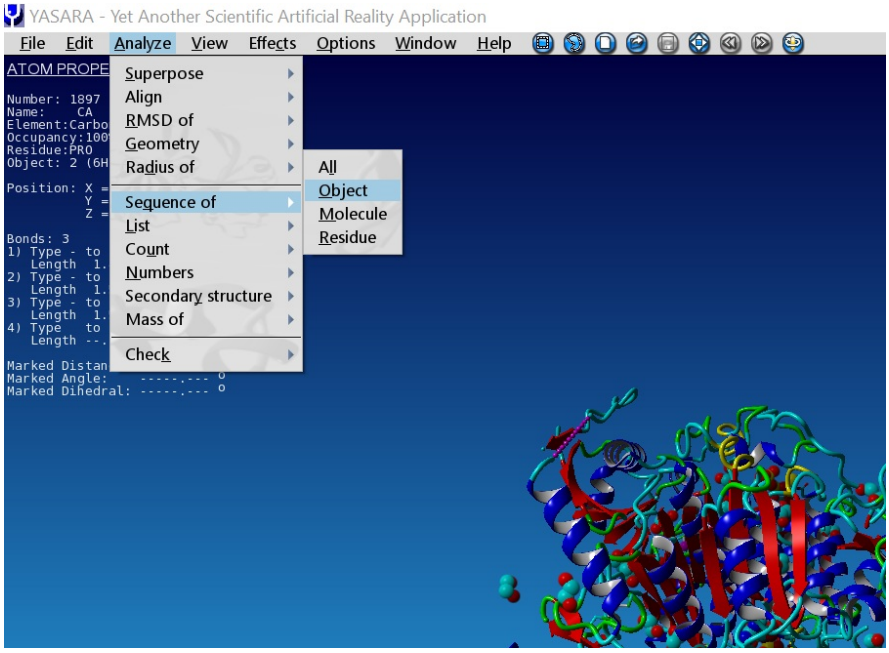
```
>sp|Q9GZZ9|UBA5_HUMAN Ubiquitin-like modifier-activating enzyme 5 OS=Homo sapiens OX=9606 GN=UBA5 PE=1 SV=1
MAESVERLQQRVQELERELAQERSLQVPRSGDGGGGRVRIEKMSSEVVDSNPYSRLMALK
RMGIVSDYEKIRTFAVAIVGVGGVGSVTAEMLTRCGIGKLLLFDYDKVELANMNRLFFQP
HQAGLSKVQAAEHTLRNINPDVLFEVHNYNITTVENFQHFMDRISNGGLEEGKPVDLVLS
CVDNFEARMTINTACNELGQTWMESGVSENAVSGHIQLIIPGESACFACAPPLVVAANID
EKTLKREGVCAASLPTTMGVVAGILVQNVLKFLLNFGTVSFYLGYNAMQDFFPTMSMKPN
PQCDDRNCRKQQEEYKKKVAALPKQEVIQEEEEIIHEDNEWGIELVSEVSEEELKNFSGP
VPDLPEGITVAYTIPKKQEDSVTELTVEDSGESLEDLMAKMKNM
```

The FASTA sequence.

Obtain Sequence from a PDB file 1m

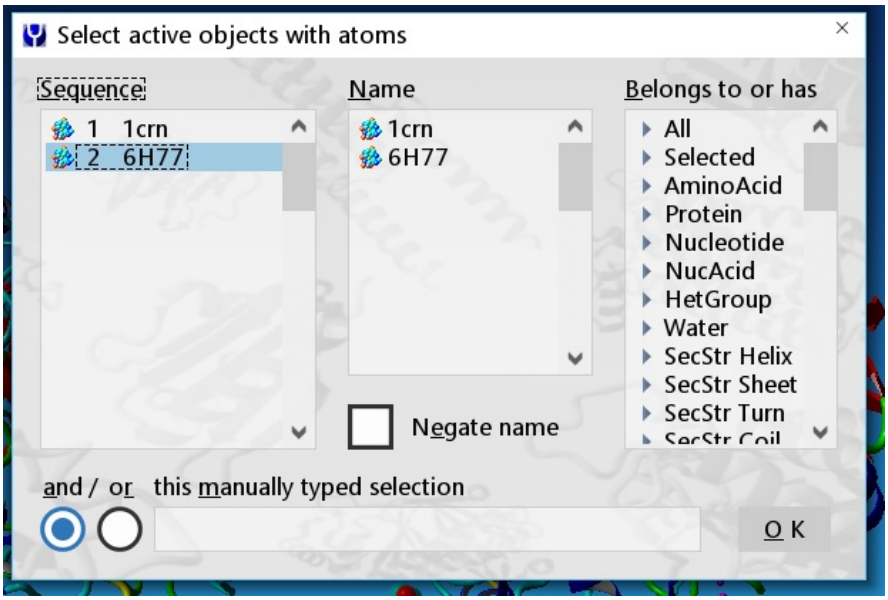
3 Extracting sequence from object on YASARA

Open your structure of choice on YASARA. Once the object is loaded, click the 'Analyze' tab at the top of the page, then 'Sequence of', then 'Object'.



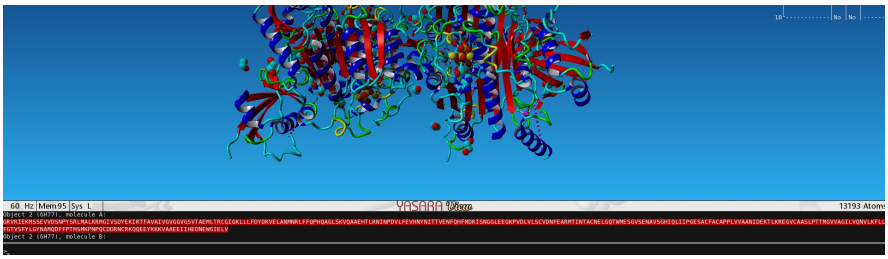
Click Analyze>>Sequence of>>Object

3.1 When the following window pops up, be sure to select the correct object. The following objects have their PDB ID's listed. The 6H77 selected is the PDB ID for the model of UBA5 being analyzed in this **example**.



Click the correct object according to its PDB ID.

3.2 The sequences should come up in the command bar at the bottom. Highlight your sequence. It will appear red once you highlight it. Press 'Cntrl C' (PC) or 'Command C' (Mac) to copy this sequence.



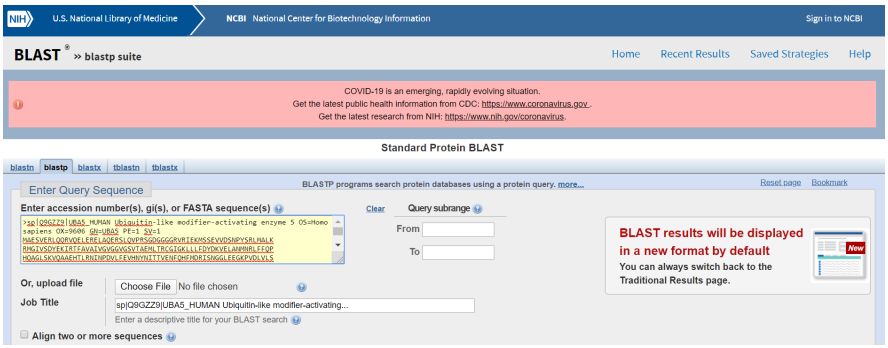
Sequence from PDB structure in the YASARA command line.

BLAST your Protein

4 Use the NCBI BLAST page to BLAST your sequence. 3m

BLAST = Basic Local Alignment Search Tool; it will search various sequence data bases to find matche to your sequences and those that are similar, which may be useful for finding proteins with similar structure or function.

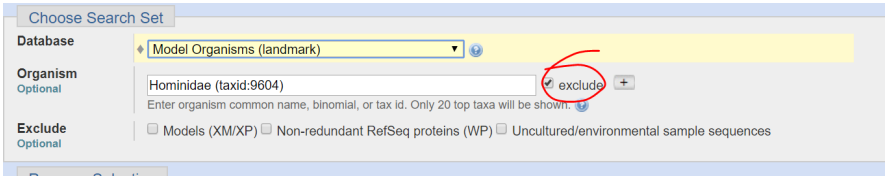
4.1 **Paste the seqeunce** you've copied from either UNIPROT or YASARA into the box under the 'Enter Query Sequence' tab.



The exame sequence is pasted and highlighted yellow in the box.

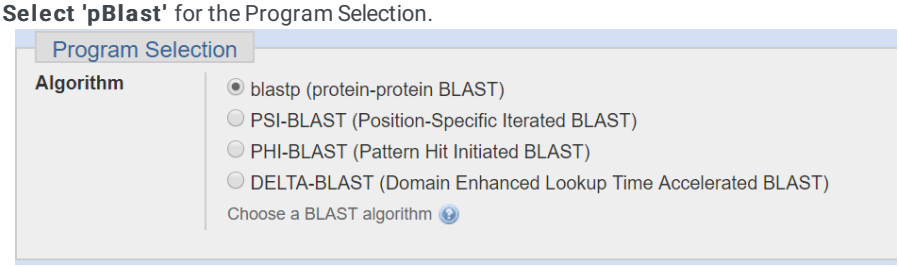
4.2 Sometimes it is useful to narrow your search by restricting the database or excluding organisms. THIS IS OPTIONAL!
The example below shows how to restrict the data base to model organisms and exclude human sequences, however this is not required for sequence alignment to work.

Select the 'Model Organism (landmark)' in the dropdown menu under 'Choose Search Set'.
Exclude Hominidae (taxid:9604) from the search. Be sure to **check the exclude box**!



This is what your 'Choose Search Set' could look like.

4.3 **Select 'pBlast'** for the Program Selection.



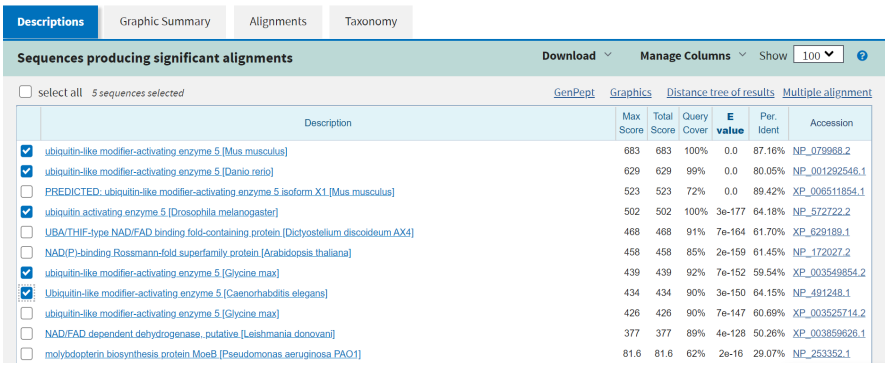
This is what your program selection should look like.

blastp is the most general search but also least adventurous search tool. PSI- or PHI-BLAST can return more hits by changing the algorithm and the weighting of sequence variation. While useful for finding distantly related proteins, it can result in some very different sequences being included. Use these with caution and experience.

4.4 **Press 'Blast'**
It may take a few minutes to a few hours before you see results.

4.5 Once the results have appeared, you can se the check boxes to the left to **select sequences** of your protein found in different species. Select options that have a query cover of 50% and above or so. Use

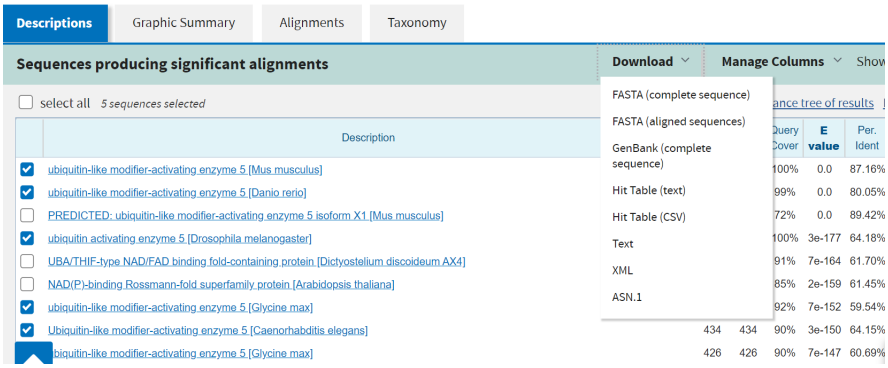
proteins with less query coverage or less than 50% sequence identity with caution



Use the check boxes to the left to select your orthologs.

Proteins that serve the same function but are in different species are called orthologs. Some of the options available are only predicted to exist indicated as PREDICTED or hypothetical protein. Avoid these as these are not yet confirmed to exist.

4.6 Once you have your desired sequences chosen, **press 'Download' >> FASTA (Complete Sequence)**



The download tab dropdown should provide the option to download the complete FASTA sequence.

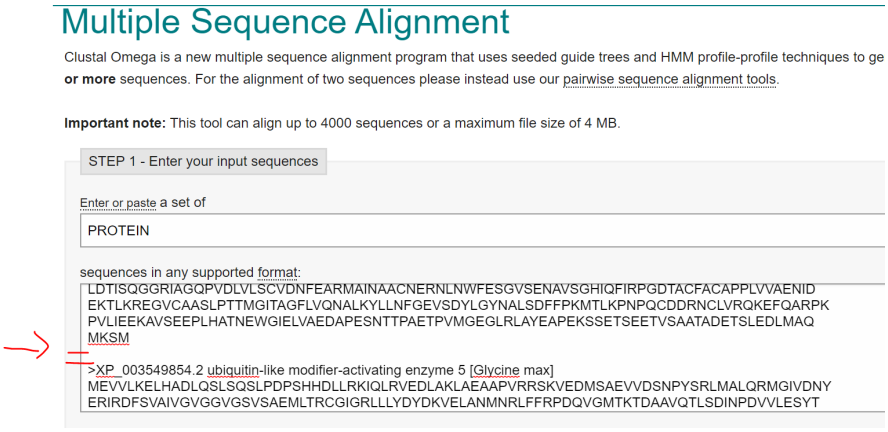
4.7 These sequences should appear in your downloads. They will open with Notepad for PC users. **Copy all the text.**

Align the sequences with Clustal Omega

5 Use Clustal Omega to align the sequences. **Click** here for the ClustalOmega homepage

5.1 **Paste** your BLAST result sequences into the 'sequences in any supported format:' box

5.2 **IMPORTANT**! The formatting for the sequences is sensitive. At the end of every sequence, **press the 'Enter' key** to make a space between each sequence.



Be sure to press enter at the end of every sequence.

5.3 **Select 'ClustalW with character counts'** for the output file.

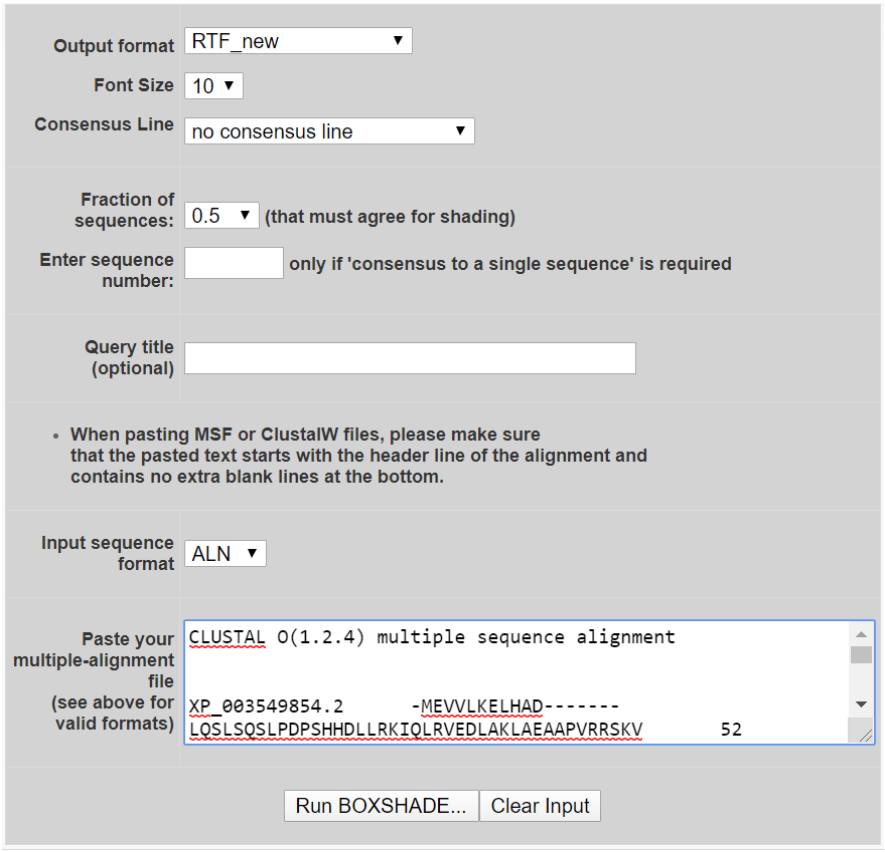
5.4 Click **'Submit'**. It may take a minute for it to produce results. Your results should look like those below.

5.5 **Copy** from the word 'Clustal' at the top of the alignment down the last character in the alignment.

Shade the alignment using Boxshade

6 Use BoxShade to better see the similarities and differences across the orthologs. **Click** here to open the BoxShade home page.

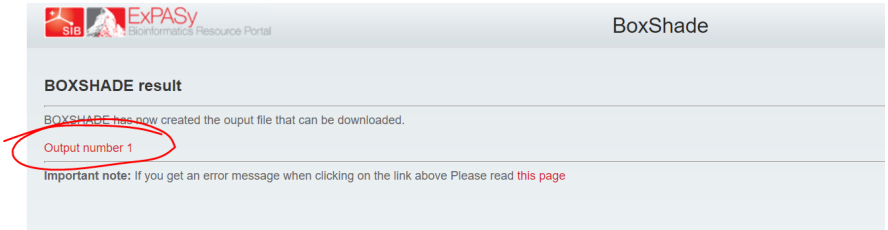
6.1 **Select 'RTF_new'** for 'Output format' and **'ALN'** for 'input sequence format'. **Paste your ClustalOmega results** into the box at the bottom.



Your input for BoxShade should look like this once you've selected the correct parameters.

6.2 **Press 'Run BOXSHADE...'** . It may take a couple of minutes.

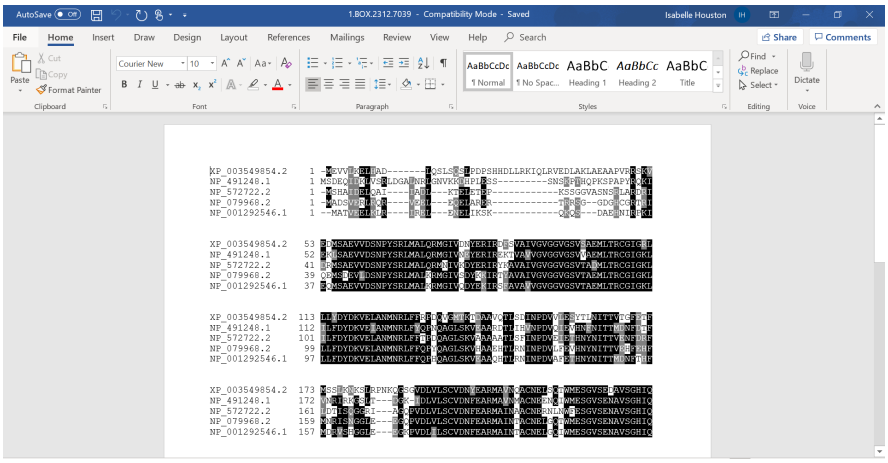
6.3 When it is done, you will get an output window. **Press on the 'Output number 1'** link provided. This will download the BoxShade file. It can be opened with Microsoft Word.



Press Output number 1

6.4 **Analyze** your results in Word.

- Black highlight means perfect conservation
- Grey highlight mean an amino acid or a position with some conservation
- No highlight means no conservation
- A dash means a gap or missing amino acid in that sequence



Final results should look something like this.