**COVID-19: On the way to a vaccine**

# Part VI - Determining the similarity of Coronavirus S (spike) glycoproteins

Learning Objective: This worksheet engages students in comparing the protein sequences of coronavirus spike proteins to understand how the SARS-CoV-2 specifically infects humans.

The S (spike) glycoprotein is a viral transmembrane protein that serves to attach the virus to the human host cell by interacting with the ACE-2 (angiotensin converting enzyme 2) of human epithelial cells. It is a key target for vaccines, therapeutic antibodies, and diagnostics. The full length S protein includes 1273 amino acids (GenBank: QIA20044.1). It forms a trimeric assembly (Figure 1) that was shown to interact with the ACE-2 protein on the cell surface through its receptor binding domain (RBD) (Gallagher and Buchmeier, 2001; Yan *et al.,* 2020). The structure of a complex of the RBD of SARS-CoV-2 S protein and the ACE-2 (angiotensin converting enzyme 2) of human epithelial cells was recently determined by Yan *et al*. (2020) and compared to a similar complex of the SARS coronavirus.

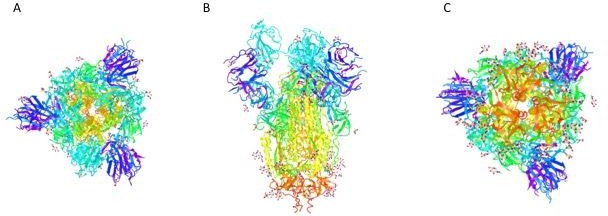


Figure 1. The trimeric structure of the SARS-CoV-2 S glycoprotein (PDB ID: 6VSB) rendered using iCn3D. Each monomeric unit has been colored using Spectrum coloring scheme, with the N-terminus colored purple and the C-terminus colored red. A) Image looking down on the protein from N-terminal end; B) side view of the protein; and C) image looking down on the protein from C-terminus end.

**Finding similarity among other coronavirus proteins**

*Compiling a list of similar sequences*:

Box 1: What is BLASTp?

The BLASTp program takes a sequence of amino acids and compares this sequence to the existing database of millions of sequences to find a match. In simple terms, the BLAST program uses an algorithm that searches ‘words’ of short amino acid sequences against the database. Matches are scored based on how similar the physicochemical characteristics of the corresponding amino acids are between the searched “word” and the prospective “match” word and then the search is repeated with another ‘word’. In addition to finding sequences with similarity, the BLAST program will provide the alignment between the search and the match and a score that reflects the quality of the match.

Goal: In this section, you will collect sequences that are similar to the RBD sequence from Yan *et al.,* 2020.

B LASTp

To find and compare sequences, you will use BLASTp.

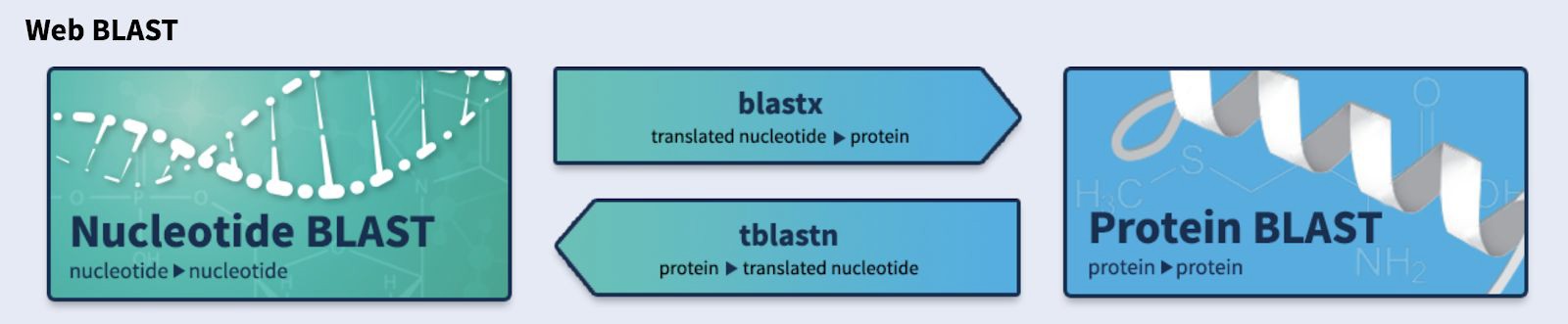
1.  Go to the NCBI BLAST website (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) You should see a page with the following search options Figure 2.

Figure 2. BLAST website.

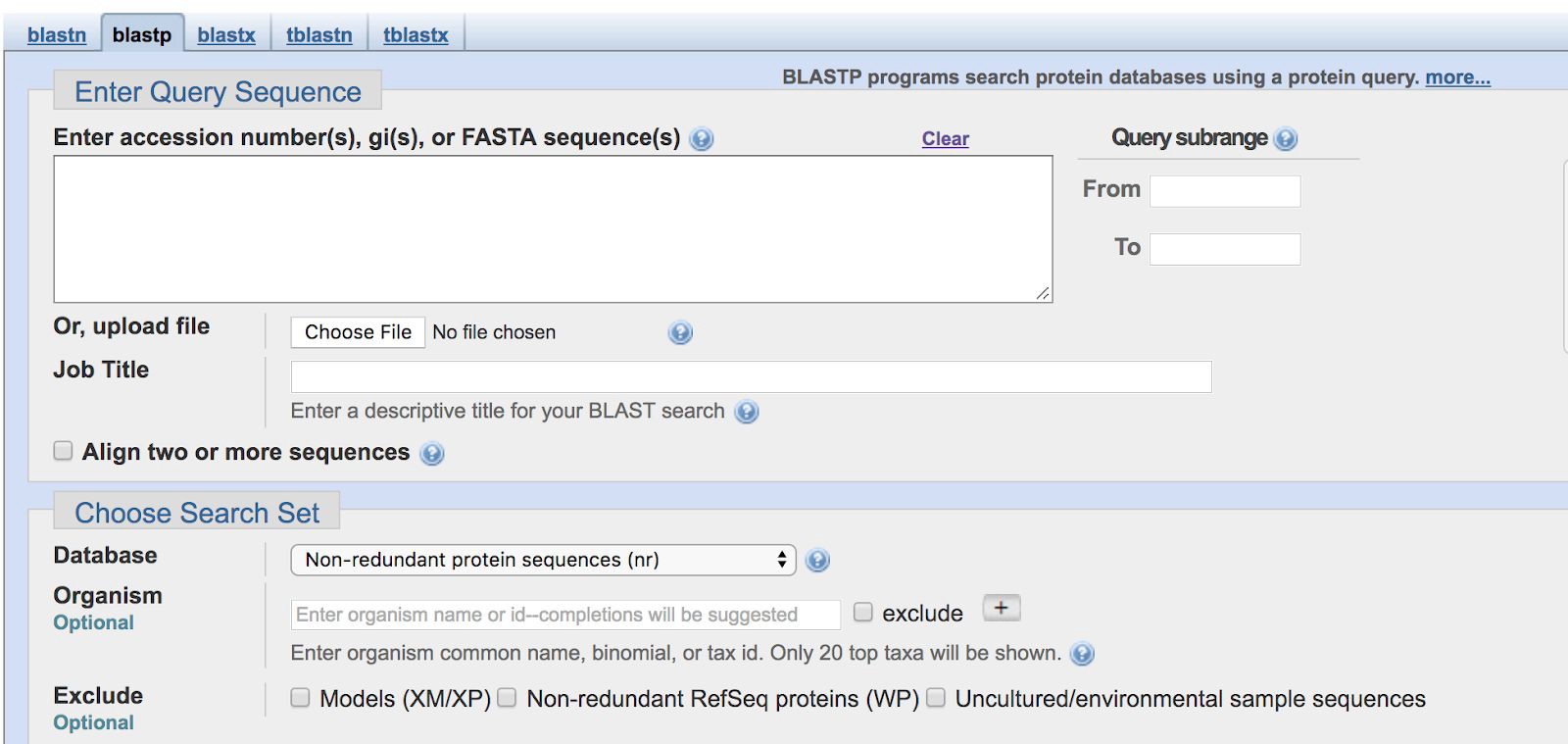
1. Click on protein BLAST (BLASTp) to do a protein-to-protein BLAST search. You should now see the BLASTp search window (Figure 3).

Figure 3. The query sequence window of BLASTp.

1. Enter “6M17\_E” into the “Enter accession number(s), gi(s), or FASTA sequence(s)” box. *This is the PDB identifier and chain identifier for the SARS-CoV-2 S glycoprotein RBD in the Spike protein-ACE-2 complex structure that was used in the molecular visualization (Yan et al., 2020).*
2. Scroll down the page and click on  to begin the search.

*The BLAST algorithm is able to get the sequence of the protein from the PDB and is now running your sequence of interest against the entire database. This may take anywhere from several seconds to a couple of minutes.*

When the BLASTp search is complete, a new window will open up. The top of that window will look similar to Figure 4.

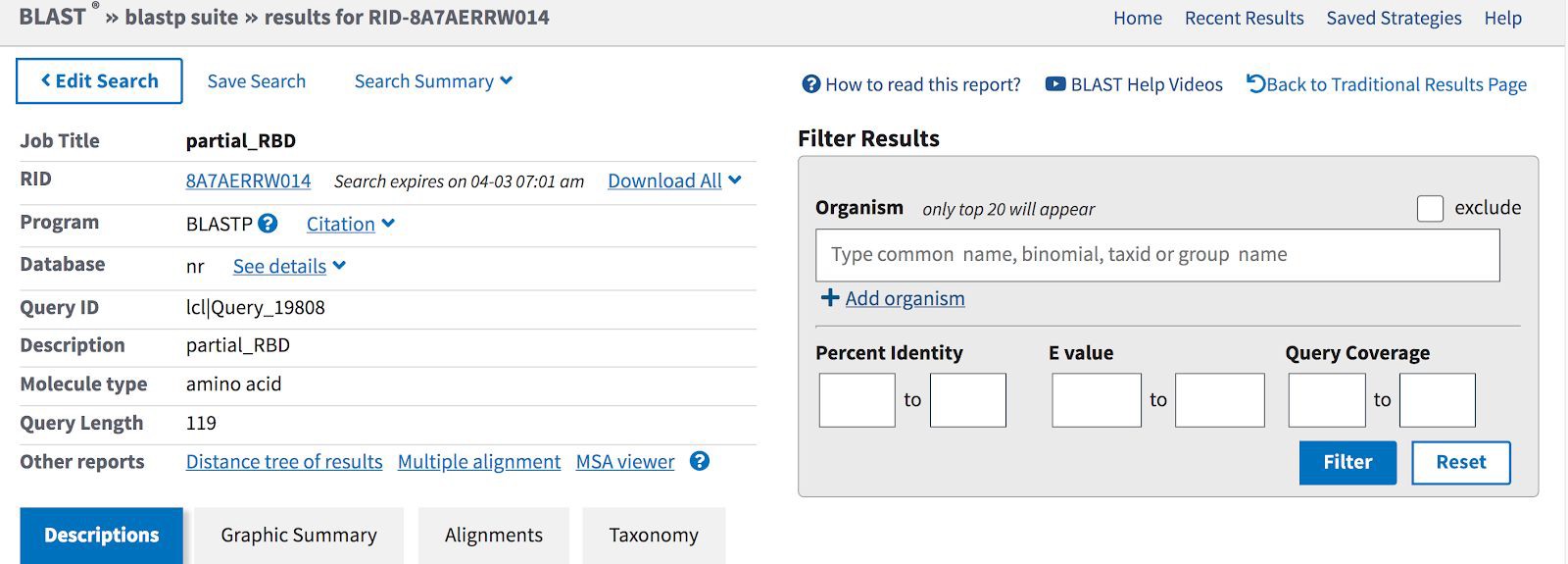


Figure 4. BLASTp search window.

1. Scroll down to find the top 100 sequences that were identified as a match to the query sequence by this search. Note that many of these sequences are identified (under Description) as “Severe acute respiratory syndrome coronavirus 2”. This is because there are many sequences from the same organism that have been submitted to the sequence database.
2. To eliminate sequences that are identical (we are looking for similar sequences) you will filter the identical sequences from the results. Go to the top of the BLASTp results page and select “Edit Search”.
3. Your 6M17\_E sequence should still be visible in the “Enter accession number(s), gi(s) or FASTA sequence(s)” box (see Figure 3).
   1. Below the sequence entry, there is a box labeled “Organism”. In this box, type “coronavirus group 2”. *As you type, some options may pop up. Select this one or completely type in the information.*
   2. To the right of the box is a button for ‘exclude’. Click on that button and a checkmark should appear. *What you are doing is eliminating the database sequences that are identified as “Coronavirus group 2”.*
   3. Repeat the BLAST search. The resulting page should have a section that looks like Figure 5.

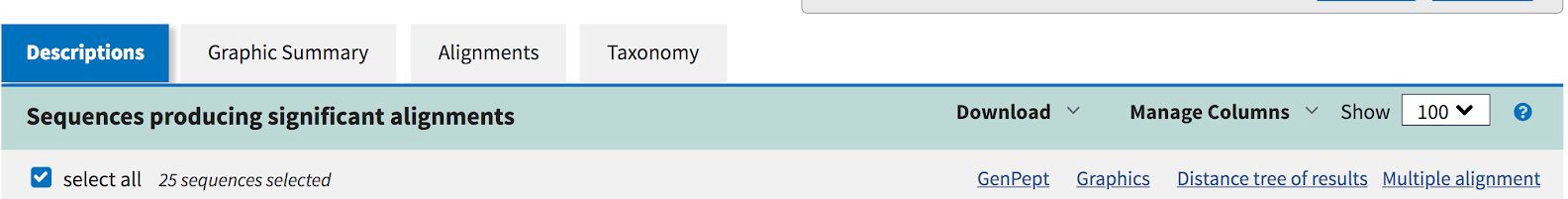


Figure 5. BLASTp results file.

1. On the left hand side there is a checked “select all” button. Click that checkmark off.
2. Scroll down in the window and you will see the descriptions of the identified sequences.
   1. Select the accession identified as “SARS CoV S protein [Expression vector pHCM-SARS-S (FFM-1)]”. This is the sequence of the S protein from the 2002 SARS coronavirus outbreak.
   2. The remaining accession numbers identify sequences from Bat (and other) coronaviruses. Select five of these sequences. *You can select any sequences that you find interesting for a total of five sequences.*
   3. Next to the select all button (which should not be selected), you should now see “6 sequences selected”.
   4. Click Download (Figure 5) and select “FASTA (Aligned sequences)”. This will create a .txt identified as “seqdump.txt”. *If you have done this a number of times, the file name may be “seqdump (#).txt”.*
   5. Download the 6M17\_E sequence by returning to the top of the results page (Figure 6), find the Query ID and click on 6M17\_E. A page will open with the sequence of 6M17\_E. Click on FASTA, copy the FASTA formatted text and paste into your text document.

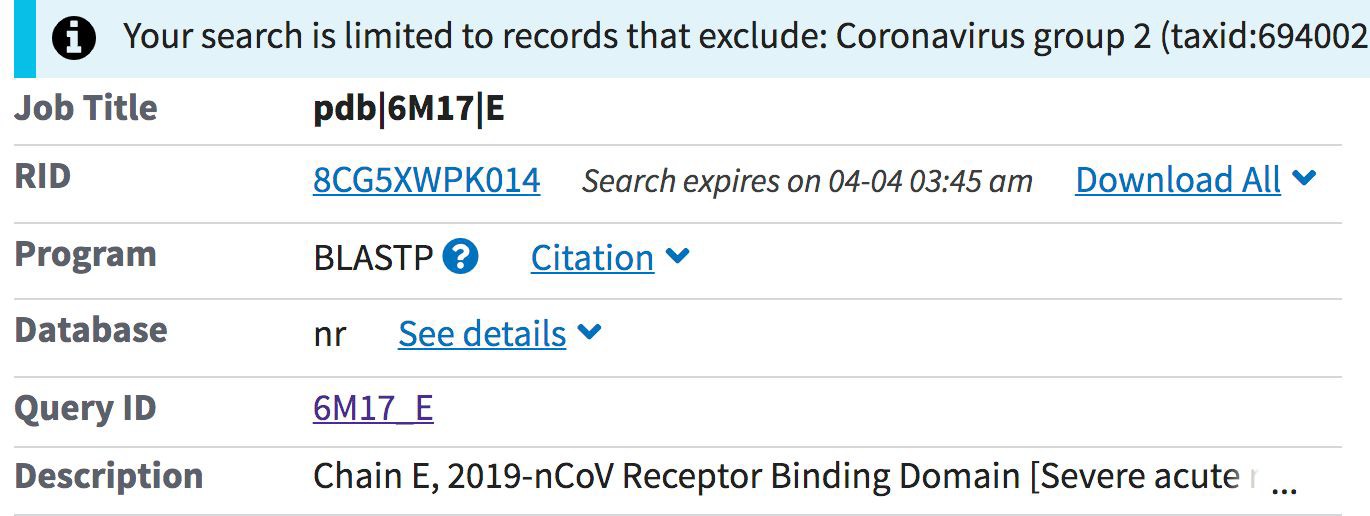


Figure 6. Finding Query ID on the BLASTp results page.

* 1. Create a .txt document that has the 6M17\_E and the six sequences that were just downloaded.
  2. Save this file which will be used for the multiple sequence alignment.

Box 2: FASTA format

FASTA is a file format that is commonly used by many bioinformatics programs. The information describing the sequence is listed after a “>” sign and followed by a return. The subsequent lines represent the amino acid sequence written in the

single letter code. When multiple FASTA sequences are used as an input (as you will do below) each sequence can be followed by a return and another entry can be started with a “>” as above. There is no limit to the number of sequences that can be included in one file in this manner

**Multiple sequence alignment (MSA)**

Goal: In this section, you will align and compare the sequences that are similar to the RBD sequence from Yan *et al.,* 2020.

When you want to compare the amino acid sequences of more than two proteins or protein fragments you need to use a multiple sequence alignment program. BLASTp will only provide the alignment of two sequences at a time. The steps here will allow you to align the three sequences you identified using BLASTp at the same time for comparison.

Note: Remember, in this exercise, you are only looking at the RBD of the S glycoprotein sequence. There may be other areas in the protein sequence that are different and may suggest significant structural and functional differences..

The BLASTp program was used to identify sequences that are similar to the SARS-CoV-2 S glycoprotein RBD. Based on the reported interaction of the SARS-CoV-2 S protein RBD with the ACE-2 protein (Yan *et al*., 2020) there are several amino acids in the COVID-19 RBD that are important in this interaction. Two of these amino acids are K417 and Y453. Note that the numbers following the amino acid identity indicate the positions of these residues in the full length protein starting with its N-terminus. Since the sequence you have used to search for similar sequences using BLASTp (6M17\_E) represents only a 223 amino acid long RBD that starts at amino acid 338 in the full sequence, the locations of these two critical residues as referenced to the start of the RBD will be K99 & Y135.

Since the sequences in your FASTA file are not the full length sequences, you will need to account for this difference using Table 2 (which is only partially completed - you will fill out the remainder).

Table 2. Mapping of critical amino acid residues for SARS-CoV-2 RBD and ACE-2 within the different sequences used for different studies. This information is based on Yan *et al*. (2020), Figure 4B-D.

|  |  |  |
| --- | --- | --- |
| **Residue location within the full length SARS-CoV-2 S protein** | **Residue location within the RBD (6M17\_E)** | **Residue location within the full length ACE-2** |

|  |  |  |
| --- | --- | --- |
| Q474 | Q156 | Q24 |
| K417 | K99 | D30 |
| Y453 | Y135 | H34 |
| N501 | N183 | Y41 |
| F486 | F168 | M85 |
| T500 | T182 | R357 |
| N501 | N183 | K353 |

CLUSTAL OMEGA alignment of the three sequences identified in BLASTp:

1. Open up CLUSTAL Omega (<https://www.ebi.ac.uk/Tools/msa/clustalo/>).

You should see a page that looks like Figure 7.

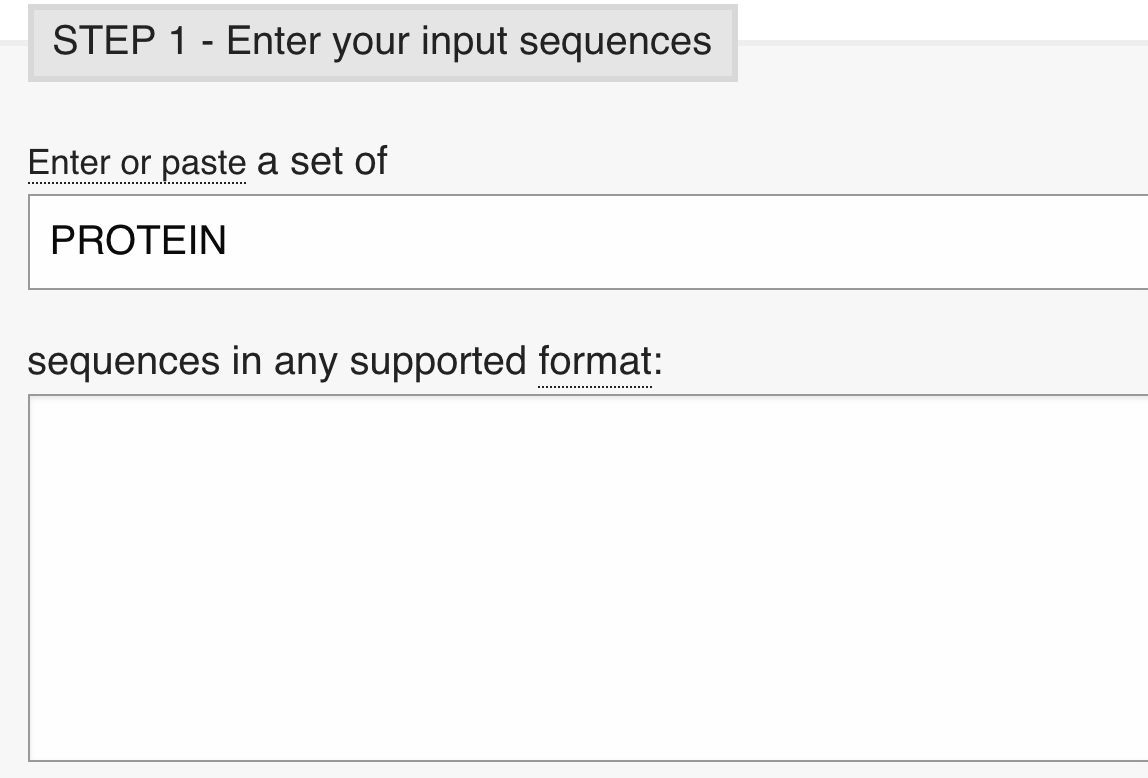


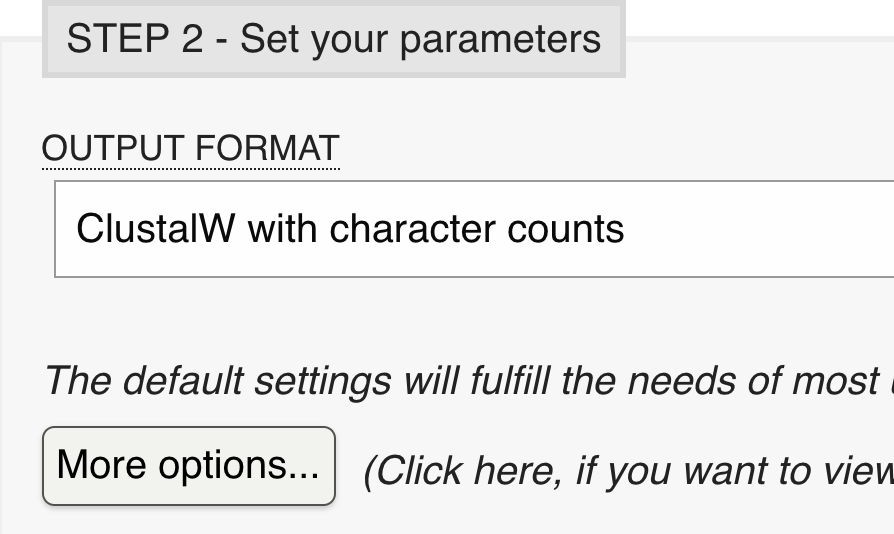
Figure 7. CLUSTAL Omega opening window.

1. Copy and paste the FASTA sequences from your .txt file into the sequences window on Clustal Omega.

*You can perform this multiple sequence alignment with many sequences that are either DNA or protein. The default (shown above the sequence window) is protein, so you do not need to change this here. You also do not need to alter any settings. The default settings will be sufficient for this approach.*

1. Scroll down to see “OUTPUT FORMAT” and click on “More options…” (Figure 8a).
2. In the expanded “OUTPUT FORMAT” window (Figure 8b), select “ORDER” and pull down “input”. *This will create the output with your first sequence (the RBD) at the top of the multiple sequence alignment.*

a.



b.

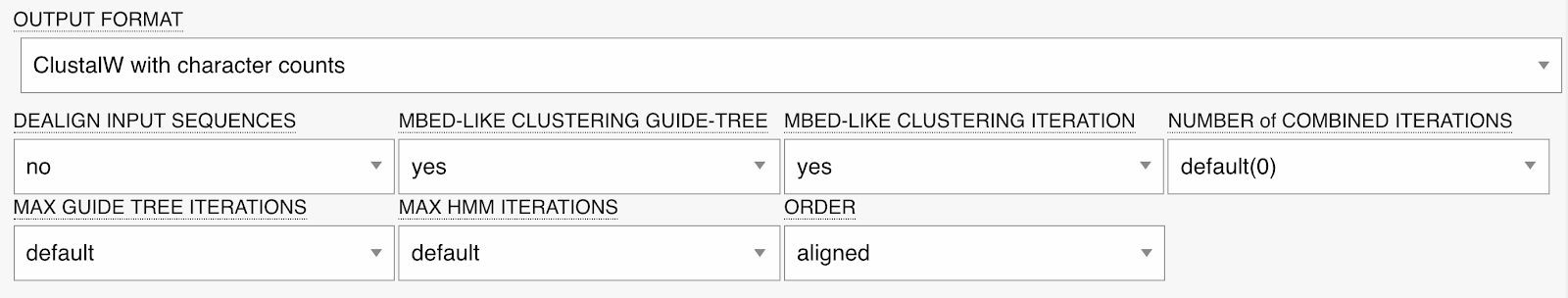


Figure 8. ClustalOmega Output format window.

1. Scroll further down and click “SUBMIT” to run the program. *The program will now run your alignment. The resulting window will provide you with the multiple sequence alignment with the heading “CLUSTAL O(1.2.4) multiple sequence alignment”*.
2. Copy the sequences and paste into a separate text file.

*If when you paste the sequences into the the text file, the alignment does not look aligned, highlight the text and change the font to Courier. This is a font in which all of the letters take up the same amount of horizontal space on a line (e.g., an m and an i take up the same space). This will make the data appear more ordered. You may also need to reduce the font size to 8 or smaller to fit in a line.*

1. Identify the sequences in the 6M17\_E from Table 2 (above). Remember than when you ran the ClustalOmega program, you set the Output format to list the aligned sequences in the Input order, so the RBD sequence should be the top line.

*Use the symbol information in Box 3 to provide some guidance rather than just counting to find the right amino acid.*

Box 3. Symbols in the alignment.

The symbols below the aligned sequences provide information about the alignment.

\* identical amino acid in all aligned sequences

: amino acids with strongly similar biochemical properties

. amino acids with weakly similar biochemical properties Space no amino acid conservation

* 1. Identify the amino acids in the aligned sequences corresponding to the amino acids that interact with ACE-2 from the Yan *et al*., (2020) paper (see Table 2).

# Questions

Based on the information in Box, use the symbols under the aligned sequences to answer the following two questions.

1. Are the seven sequences identical, similar, or unrelated over the RBD? Briefly justify your answer based on your analysis of the primary structure of these domains.
2. How conserved are the amino acids implicated in the interaction between

SARS-CoV-2 (the RBD sequence here) and ACE-2 proteins? Include in your answer a short discussion on the similarities and differences in the physicochemical properties of the amino acids at these six positions.

1. How might the differences in the SARS-CoV-2 S protein sequences identified in this exercise indicate a difference in the attachment of the virus to the host cell?