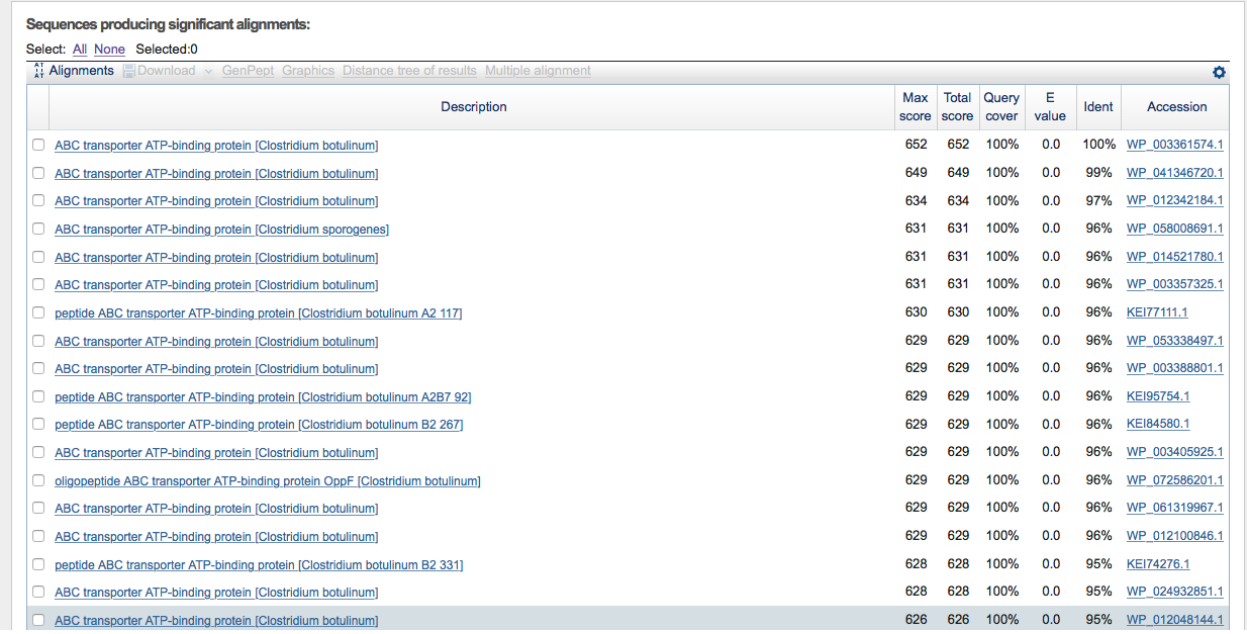
What do I do if my blast searches seem to have all the top hits from the same genus or species?

If the bacterial species you are using to annotate is clinically significant or of great research interest, you may find that when you perform blast searches (particularly in nr) that you seemingly only get hits that are different strains or isolates of the same species. This obviously doesn't give you much information about how well conserved the protein on which you are working is compared to proteins in other genera. There is a method to modify blast to let you exclude such hits from your searches.

I will use a gene from *Clostridium botulinum* as an example to illustrate this using the protein sequence of the gene with the locus tag CLJ_B3418. Figure 1 shows the top nr blast hits for this protein. You can easily see that all of the hits but one are from *Clostridium botulinum* with very high levels of coverage and identities. They are essentially all the same protein from different isolates of *Clostridium botulinum*.



Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
ABC transporter ATP-binding protein [Clostridium botulinum]	652	652	100%	0.0	100%	WP_003361574.1
ABC transporter ATP-binding protein [Clostridium botulinum]	649	649	100%	0.0	99%	WP_041346720.1
ABC transporter ATP-binding protein [Clostridium botulinum]	634	634	100%	0.0	97%	WP_012342184.1
ABC transporter ATP-binding protein [Clostridium sporogenes]	631	631	100%	0.0	96%	WP_058008691.1
ABC transporter ATP-binding protein [Clostridium botulinum]	631	631	100%	0.0	96%	WP_014521780.1
ABC transporter ATP-binding protein [Clostridium botulinum]	631	631	100%	0.0	96%	WP_003357325.1
peptide ABC transporter ATP-binding protein [Clostridium botulinum A2 117]	630	630	100%	0.0	96%	KEI77111.1
ABC transporter ATP-binding protein [Clostridium botulinum]	629	629	100%	0.0	96%	WP_053338497.1
ABC transporter ATP-binding protein [Clostridium botulinum]	629	629	100%	0.0	96%	WP_003388801.1
peptide ABC transporter ATP-binding protein [Clostridium botulinum A2B7 92]	629	629	100%	0.0	96%	KEI95754.1
peptide ABC transporter ATP-binding protein [Clostridium botulinum B2 267]	629	629	100%	0.0	96%	KEI84580.1
ABC transporter ATP-binding protein [Clostridium botulinum]	629	629	100%	0.0	96%	WP_003405925.1
oligopeptide ABC transporter ATP-binding protein OppF [Clostridium botulinum]	629	629	100%	0.0	96%	WP_072586201.1
ABC transporter ATP-binding protein [Clostridium botulinum]	629	629	100%	0.0	96%	WP_061319967.1
ABC transporter ATP-binding protein [Clostridium botulinum]	629	629	100%	0.0	96%	WP_012100846.1
peptide ABC transporter ATP-binding protein [Clostridium botulinum B2 331]	628	628	100%	0.0	95%	KEI74276.1
ABC transporter ATP-binding protein [Clostridium botulinum]	628	628	100%	0.0	95%	WP_024932851.1
ABC transporter ATP-binding protein [Clostridium botulinum]	626	626	100%	0.0	95%	WP_012048144.1

Figure 1. The blast results using a non-filtered nr blast search for CLJ_B3418.

The blast search can be set up slightly differently to prevent this problem from occurring. As noted in figure 2, we can set the search up to exclude, in this case, the taxid: 1485 (Clostridium). The taxid number stands for the NCBI Taxonomy ID number. By excluding the taxid number 1485, all blast hits in that taxonomic classification will not be included. To do this we type the genus name Clostridium in the Organism textbox below the sequence input box. As you type a pulldown menu of options will appear which you can subsequently just click on to

select (see highlighted menu item in Figure 2). The simply click the Exclude checkbox next to the organism name and then select blast.

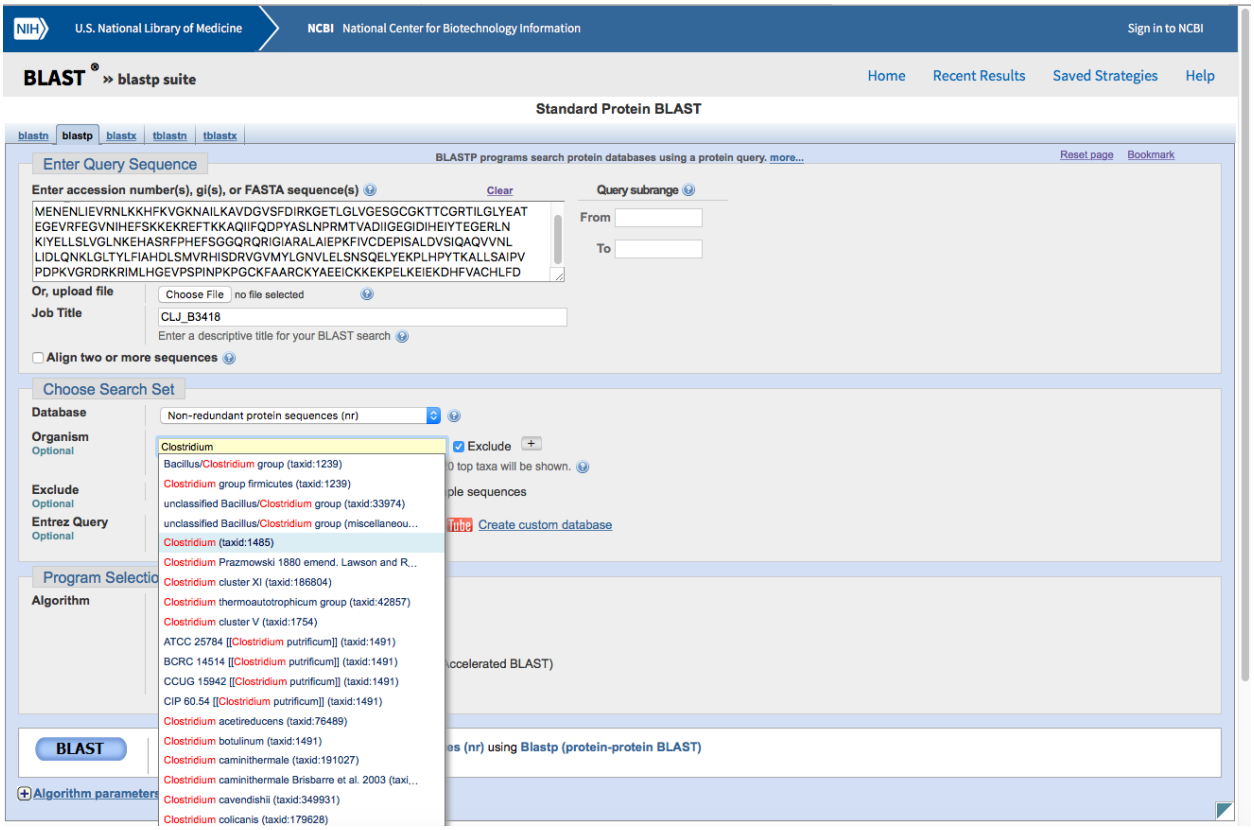
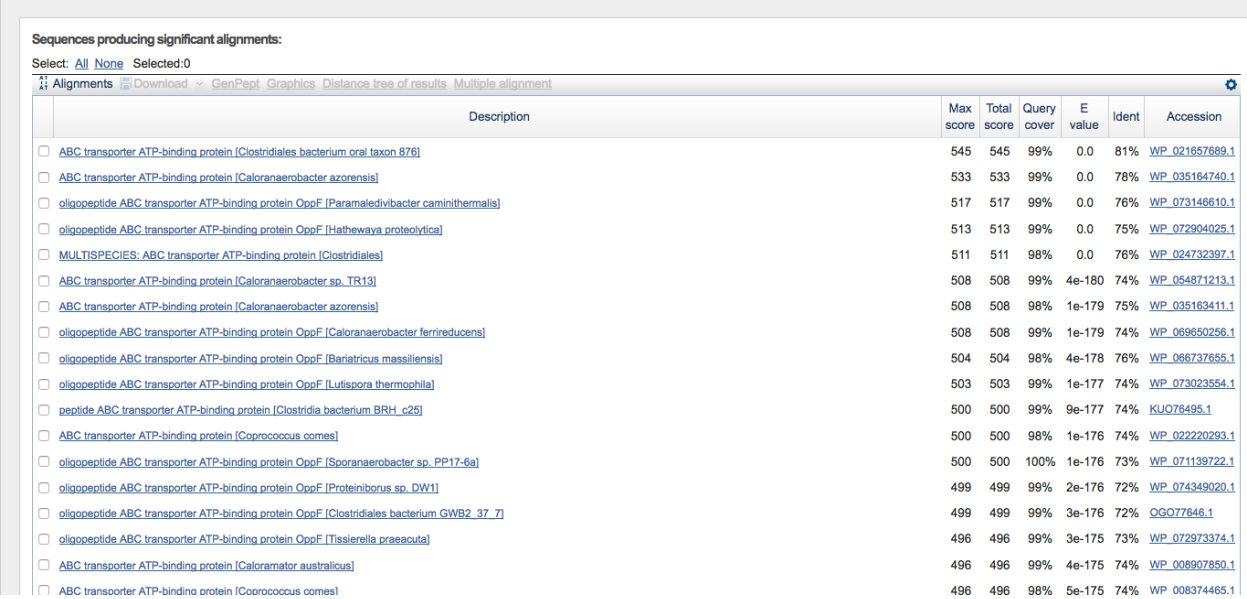


Figure 2. Setting up a blast search to exclude the *Costridium* taxid: 1485.

Figure 3 shows the results of the blast result for the same protein AFTER excluding the *Clostridium* taxid 1485. Note the different names appearing in the search results. However, also note that the top hit is no longer the one that matches the protein under investigation in the species you are working on. Thus you would take the FIRST nr hit as the top hit in this case instead of skipping over the first one. ***Note also that if you use this blast result to select sequences for the T-Coffee alignment that you will subsequently do in the Sequence Based Similarity Module, that you will need to add the FASTA formatted sequence of the protein under investigation to the top of the list before constructing the alignment.*** Students should also add a comment in their textbook of which taxid number was excluded from their search.

Experiment with different levels of exclusion (only one species) or add multiple options for exclusion (i.e., the genus) or somewhere in between (different specific species excluded by adding additional organism boxes in which to enter choices by using the + option to the right of the exclude check box to add another).



Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
ABC transporter ATP-binding protein [Clostridiales bacterium oral taxon 876]	545	545	99%	0.0	81%	WP_021657689.1
ABC transporter ATP-binding protein [Caloranaerobacter azorensis]	533	533	99%	0.0	78%	WP_035164740.1
oligopeptide ABC transporter ATP-binding protein OppF [Paramaledivibacter caminithermalis]	517	517	99%	0.0	76%	WP_073146610.1
oligopeptide ABC transporter ATP-binding protein OppF [Hathewaya proteolytica]	513	513	99%	0.0	75%	WP_072904025.1
MULTISPECIES: ABC transporter ATP-binding protein [Clostridiales]	511	511	98%	0.0	76%	WP_024732397.1
ABC transporter ATP-binding protein [Caloranaerobacter sp. TR13]	508	508	99%	4e-180	74%	WP_054871213.1
ABC transporter ATP-binding protein [Caloranaerobacter azorensis]	508	508	98%	1e-179	75%	WP_035163411.1
oligopeptide ABC transporter ATP-binding protein OppF [Caloranaerobacter ferrireducens]	508	508	99%	1e-179	74%	WP_069650256.1
oligopeptide ABC transporter ATP-binding protein OppF [Bariatricus massiliensis]	504	504	98%	4e-178	76%	WP_066737655.1
oligopeptide ABC transporter ATP-binding protein OppF [Lutispora thermophila]	503	503	99%	1e-177	74%	WP_073023554.1
peptide ABC transporter ATP-binding protein [Clostridia bacterium BRH_c25]	500	500	99%	9e-177	74%	KUO76495.1
ABC transporter ATP-binding protein [Coprococcus comes]	500	500	98%	1e-176	74%	WP_022220293.1
oligopeptide ABC transporter ATP-binding protein OppF [Sporanaerobacter sp. PP17-6a]	500	500	100%	1e-176	73%	WP_071139722.1
oligopeptide ABC transporter ATP-binding protein OppF [Proteiniborus sp. DW1]	499	499	99%	2e-176	72%	WP_074349020.1
oligopeptide ABC transporter ATP-binding protein OppF [Clostridiales bacterium GWB2_37_7]	499	499	99%	3e-176	72%	OGO77646.1
oligopeptide ABC transporter ATP-binding protein OppF [Tissierella praeacuta]	496	496	99%	3e-175	73%	WP_072973374.1
ABC transporter ATP-binding protein [Caloramator australicus]	496	496	99%	4e-175	74%	WP_008907850.1
ABC transporter ATP-binding protein [Coprococcus comes]	496	496	98%	5e-175	74%	WP_008374465.1

Figure 3. The nr blast results for CLJ_B3418 AFTER excluding the *Costridium* taxid: 1485. Note the different genus and species names of the top hits.

You can also use the NCBI Taxonomy Browser to find different levels of taxid to use in your exclusion searches, especially if it is not clear what you should choose from the pulldown menu in BLAST. The use of the Taxonomy Browser is described in general terms in the Horizontal Gene Transfer section of the project manual. Briefly, go to: https://www.ncbi.nlm.nih.gov/taxonomy and enter the name of your organism's genus in the search window (*Clostridium botulinum* is the example used below) and click on Search. A result similar to Figure 4 will display. Click on the organism hyperlink in blue, and you will be taken to the full lineage of the organism (next figure).

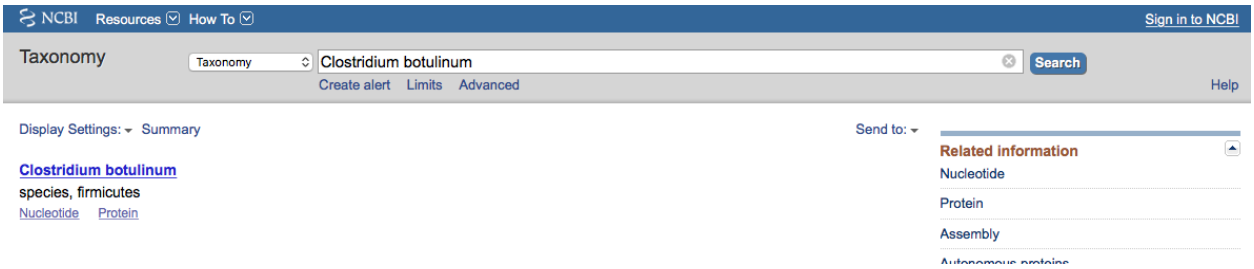


Figure 4. Results of searching for Clostridium botulinum in the NCBI Taxonomy browser.

Figure 5 below shows a portion of the C. botulinum results. In the lineage line, the last entry is the genus (Clostridum), but you can hover the cursor over any of the levels of taxonomy and see the name of the level (i.e., family, order etc.). Figure 6 shows what will display when the Clostridium hyperlink is selected.

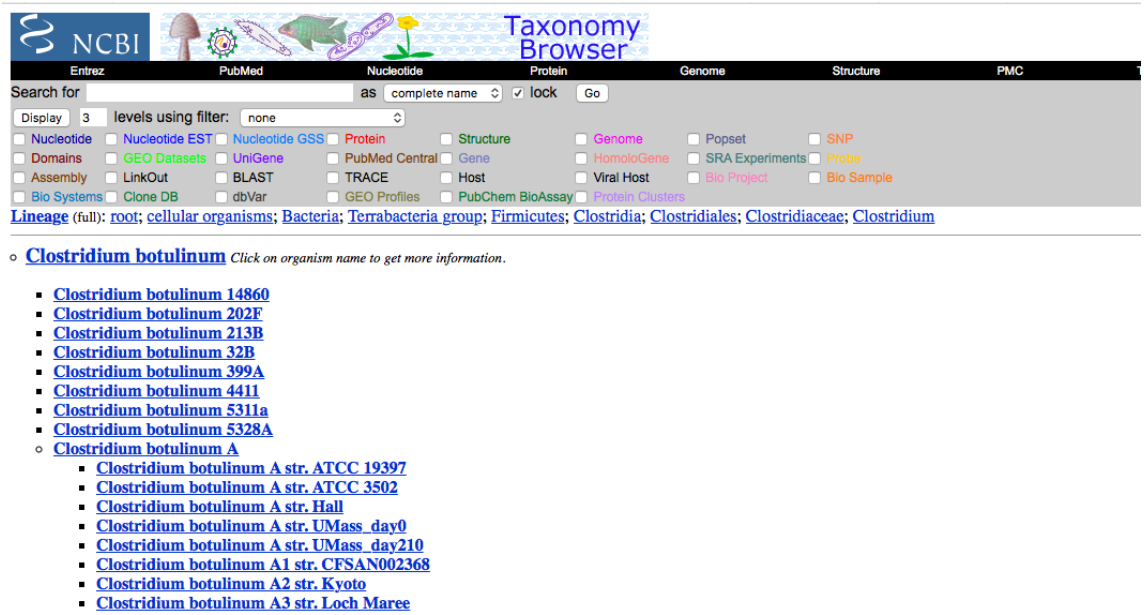


Figure 5. The Clostridium botulinum results from NCBI Taxonomy browser (not complete).

Of interest in the results from clicking on the Clostridium hyperlink displayed displayed in Figure 6 is the Taxonomy ID of 1485 (exactly the one we found by limiting the BLAST results from within the BLAST tool). You could use this information to simply type "Clostridium (taxid:1485)" – do not, however, include the quotation marks- in the organism window of the BLAST search and click exclude as before.

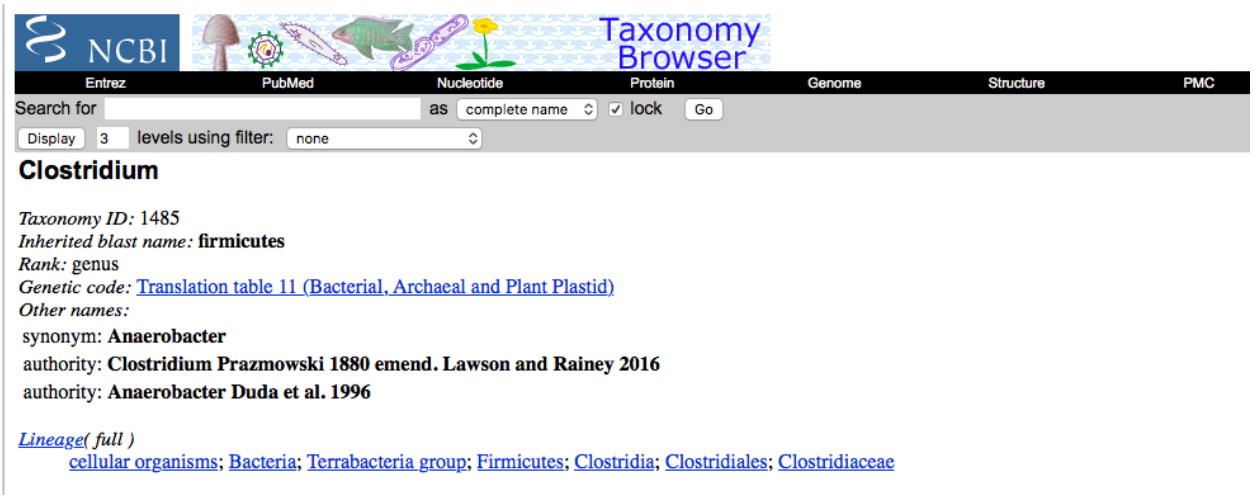


Figure 6. The Clostridium genus taxid information.

We can also go further “up” in taxonomic window to exclude more than one genus (though you should not have to do that routinely). For example, Figure 7 shows the display that would come up if we clicked on the Clostridiaceae (i.e., the family to which the genus Clostridum belongs) hyperlink instead of the Clostridium hyperlink. Different genera will appear that are part of this family. Clicking on the Clostridiaceae link from this page will result in the information shown in the Figure 8.

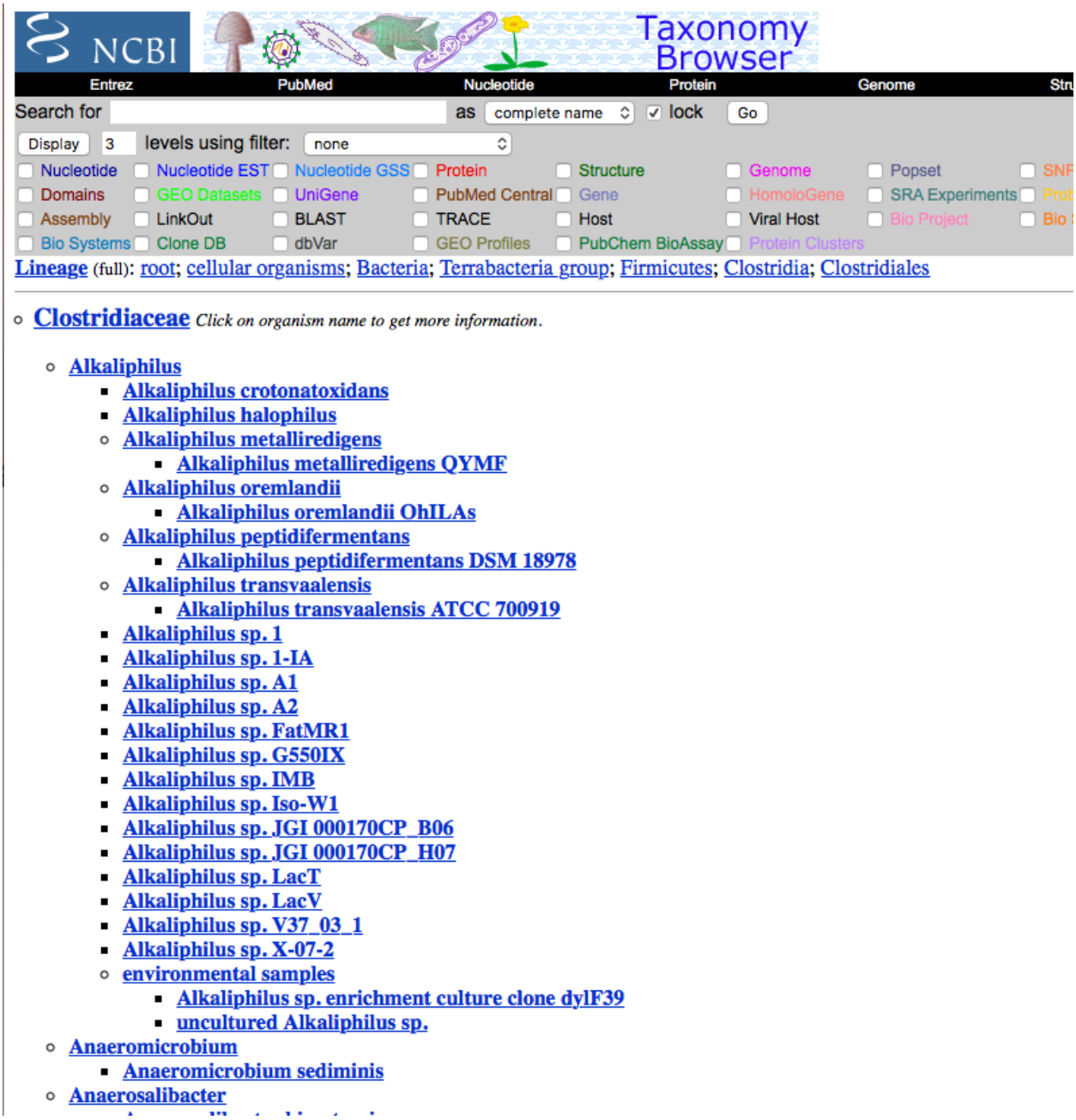


Figure 7. The display resulting from selection of the Family Clostridiaceae in the taxonomy browser.

Here we see that the Clostridiaceae family has the Taxonomy ID of 31979. To exclude this Family from the BLAST results, we would simply type in "Clostridiaceae (taxid:31979)" into the organism box in the BLAST search and click exclude. The next image will show how the autofill option will highlight once we paste in the taxid.

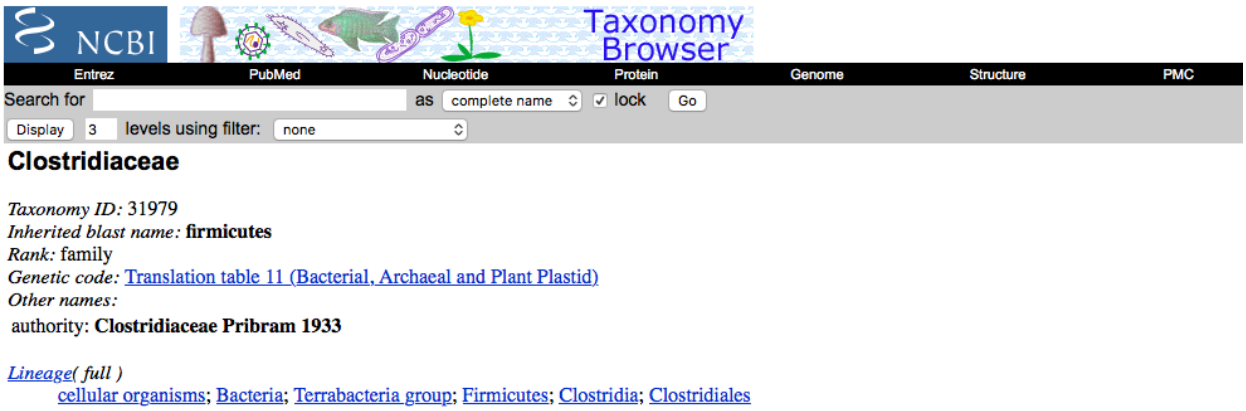


Figure 8. The taxonomy identification number of the Family Clostridiaceae.

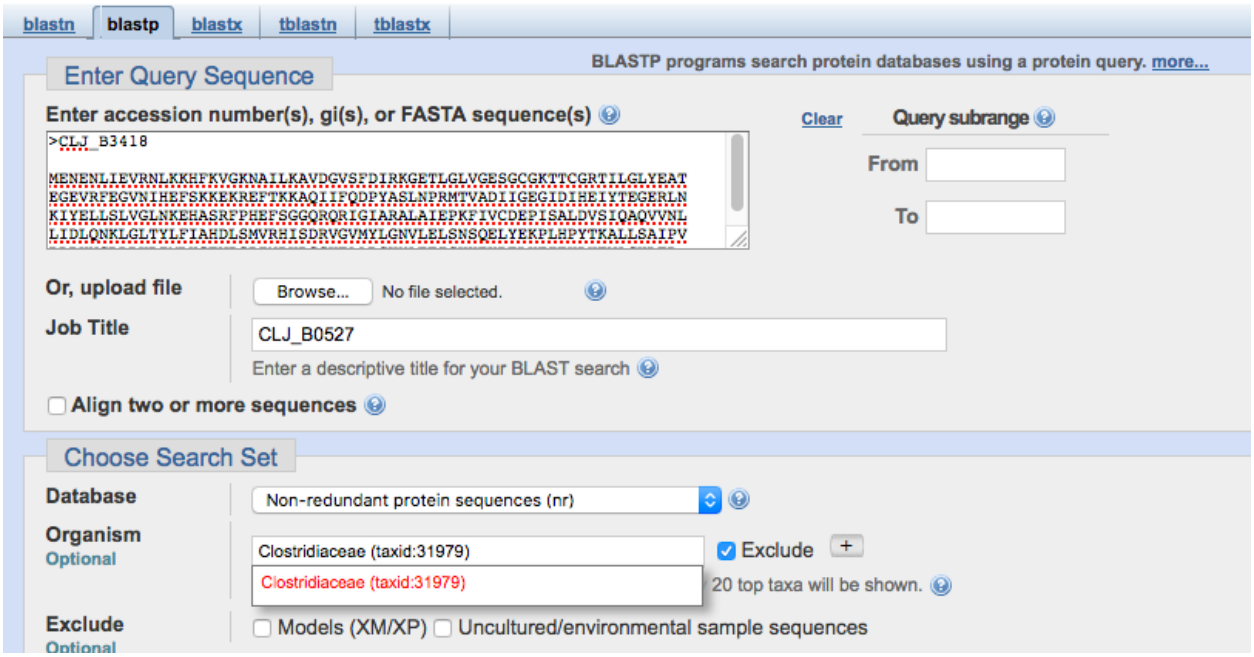
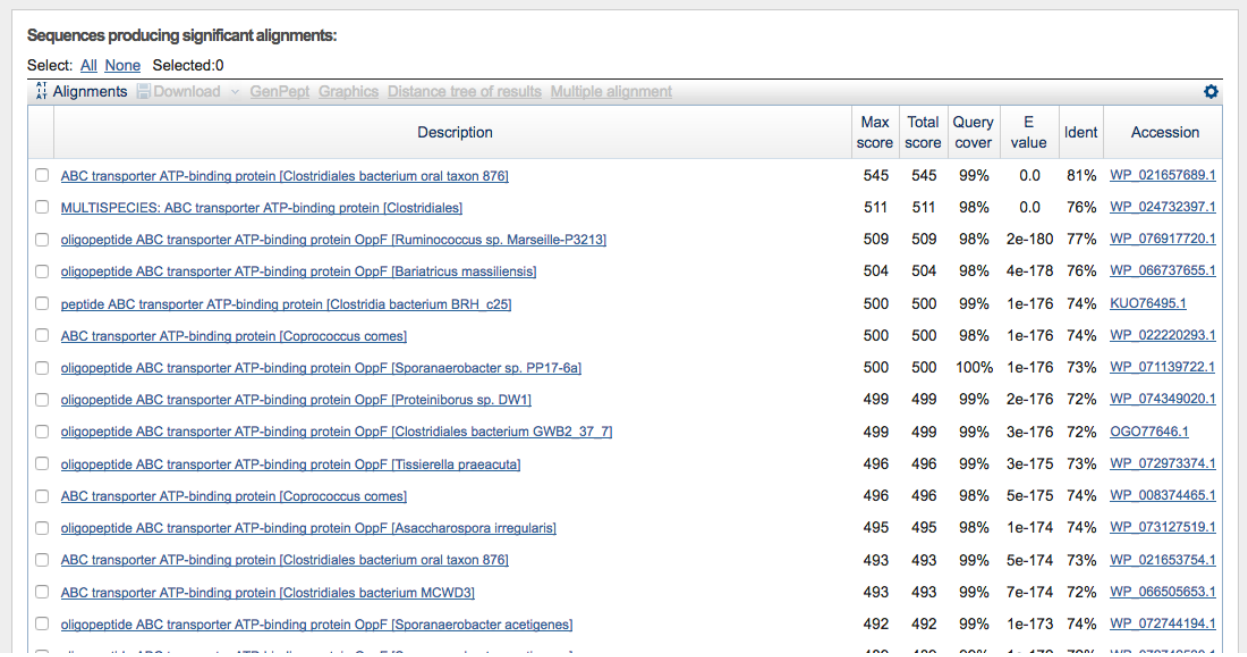


Figure 9. A BLAST search set up to exclude members of the Family Clostridiaceae from the search results.

Finally, Figure 10 shows the BLAST results from doing the exclusion at this level.



Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
ABC transporter ATP-binding protein [Clostridiales bacterium oral taxon 876]	545	545	99%	0.0	81%	WP_021657689.1
MULTISPECIES: ABC transporter ATP-binding protein [Clostridiales]	511	511	98%	0.0	76%	WP_024732397.1
oligopeptide ABC transporter ATP-binding protein OppF [Ruminococcus sp. Marseille-P3213]	509	509	98%	2e-180	77%	WP_076917720.1
oligopeptide ABC transporter ATP-binding protein OppF [Bariatricus massiliensis]	504	504	98%	4e-178	76%	WP_066737655.1
peptide ABC transporter ATP-binding protein [Clostridia bacterium BRH_c25]	500	500	99%	1e-176	74%	KUO76495.1
ABC transporter ATP-binding protein [Coprococcus comes]	500	500	98%	1e-176	74%	WP_022220293.1
oligopeptide ABC transporter ATP-binding protein OppF [Sporanaerobacter sp. PP17-6a]	500	500	100%	1e-176	73%	WP_071139722.1
oligopeptide ABC transporter ATP-binding protein OppF [Proteiniborus sp. DW1]	499	499	99%	2e-176	72%	WP_074349020.1
oligopeptide ABC transporter ATP-binding protein OppF [Clostridiales bacterium GWB2_37_7]	499	499	99%	3e-176	72%	OGO77646.1
oligopeptide ABC transporter ATP-binding protein OppF [Tissierella praeacuta]	496	496	99%	3e-175	73%	WP_072973374.1
ABC transporter ATP-binding protein [Coprococcus comes]	496	496	98%	5e-175	74%	WP_008374465.1
oligopeptide ABC transporter ATP-binding protein OppF [Asaccharospora irregularis]	495	495	98%	1e-174	74%	WP_073127519.1
ABC transporter ATP-binding protein [Clostridiales bacterium oral taxon 876]	493	493	99%	5e-174	73%	WP_021653754.1
ABC transporter ATP-binding protein [Clostridiales bacterium MCWD3]	493	493	99%	7e-174	72%	WP_066505653.1
oligopeptide ABC transporter ATP-binding protein OppF [Sporanaerobacter acetigenes]	492	492	99%	1e-173	74%	WP_072744194.1

Figure 10. BLAST results after excluding the Family Clostridiaceae from the search.