HOW TO GET USABLE AMINO ACID AWARE NUCLEOTIDE ALIGNMENTS

In Milestone 1, you are asked to carry out two searches: a search of nucleic acid databases to find very closely related sequences and an amino acid search to find more distant members of your gene family. In the practica, you were introduced to searching for and downloading amino acid sequences. When working with nucleic acid sequences, there are some additional hurdles to downloading sequences and obtaining multiple alignments. Specifically,

- in your nucleic acid searches, the matching sequence retrieved by blast is typically an entire genome. You need to extract just the protein coding sequence (CDS) in the genome that is similar to the query.

- At the same time, you will want to extract a label for the matching sequence that includes the name of its genome, as well as the amino acid sequence encoded by your sequence.

- Before proceeding, relabel your sequences using locus tags (described below). For every nucleotide sequence, the same label should also be assigned to corresponding amino acid sequence.

- Some nucleic acid sequences are located on the negative strand in the genomic DNA. Each of these sequences must be replaced with its reverse-complement.

- When aligning nucleotide sequences that encode proteins, a more accurate alignment can be obtained if the alignment is guided by the amino acid sequence to ensure that codon positions are correctly aligned. This requires that the amino acid translations of your nucleotide sequences be aligned in a pre-processing step. .

- Obtain a "codon aware" nucleotide alignment using the amino acid multiple alignment and a tool called PAL2NAL.

Below we describe each of these steps in detail.

NOMENCLATURE

As we discussed in Practicum 1, it is useful to label sequences in a data set with short labels (typically not more than 10 or 15 characters) that are unique and have some mnemonic value. Because we will be comparing gene trees to a species tree in this study, sequence labels must also include information about the genome from which the sequence was originally derived. To do this, we can take advantage of a naming convention already in use in sequencing projects: *locus tags*.

When a genome sequence is submitted to a public database, such as Genbank, it is assigned a unique identifier called a "locus tag prefix". Annotated genes in the sequence are numbered

sequentially, sometimes with gaps in the numbering system. Each gene is then assigned a unique name that comprises the prefix and the gene number. For example, the locus tag prefix of the genome sequence for *Roseburia intestinalis XB6B4* is "RO1". The locus tags of individual genes in this genome are RO1_00110, RO1_00120, RO1_00130…

Note that each genome (not each species) has a unique locus tag prefix. If multiple strains of the same species are sequenced, each one will have a different locus tag prefix.

Since locus tags are short, unique, and contain the name of the associated genome sequence, they are a good choice for naming sequences in our project. Locus tags are stored in Genbank records, so you will be able to obtain the locus tags at the same time that you download your selected sequences from the database.

EXTRACTING APPROPRIATE NUCLEIC AND AMINO ACID SEQUENCES FROM A DATABASE SEARCH

Once you have performed your BLAST search (be it blastn, blastp , or tblastx) you will have a list of genes that were found to be similar to your query. In the blast context, a database sequence is similar to the query if there is a high-scoring match between a local region in the query and a local region in the database sequence. However, in most cases, the sequence of interest is not just the subsequence in that high-scoring local region, but a longer sequence that contains it.

When searching amino acid databases, the sequences you retrieve typically correspond to a single protein. In this case, what you want is the entire sequence. You can check the sequences that look promising and then click "Download" to obtain a collection of amino acid sequences in fasta format. This is what you did in Practicum 1.

With nucleic acid searches, the high-scoring local region is, in many cases, located in an entire genome sequence, which is usually not what you want. In this project, the high-scoring local region will usually be contained in a subsequence of the genome that has been annotated as a protein coding sequences (CDS). In this case, it is the entire CDS, not the entire genome, that is of interest. Here we describe how to extract the CDS from the genome sequence retrieved by blast. You can obtain the locus tag and the amino acid sequence encoded by the CDS at the same time.

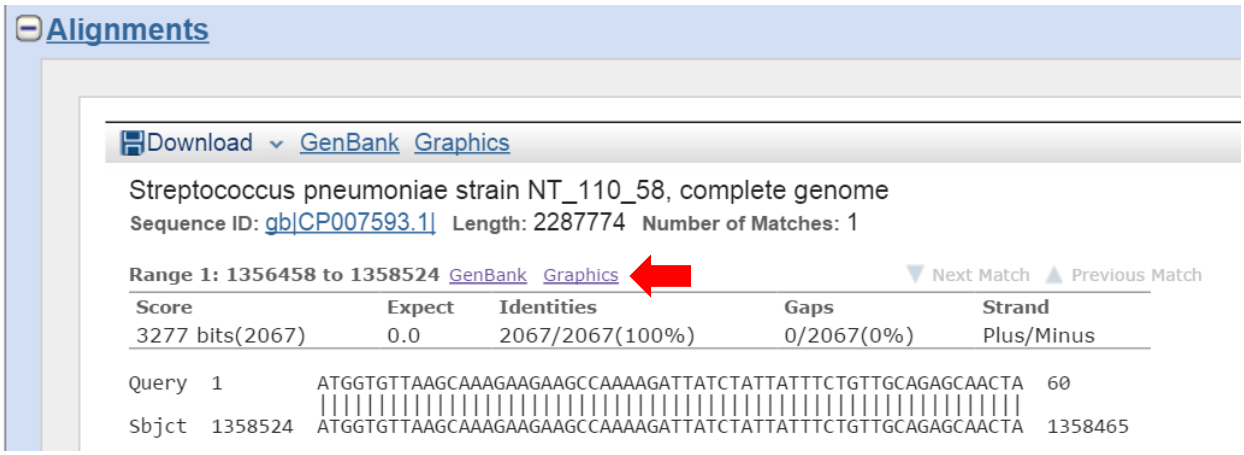
You have completed your blast searches and have identified a set of nucleic acid sequences for analysis. To recover the complete sequence of both the amino acid and nucleotide sequences of the genes on your list, perform the following steps:

1. First we will be copying sequences of amino acids and nucleotides. For this you will need two permanent files
    a. *Amino_acid_sequences.fasta*
    b. *All_nucleotide_sequences.fasta*

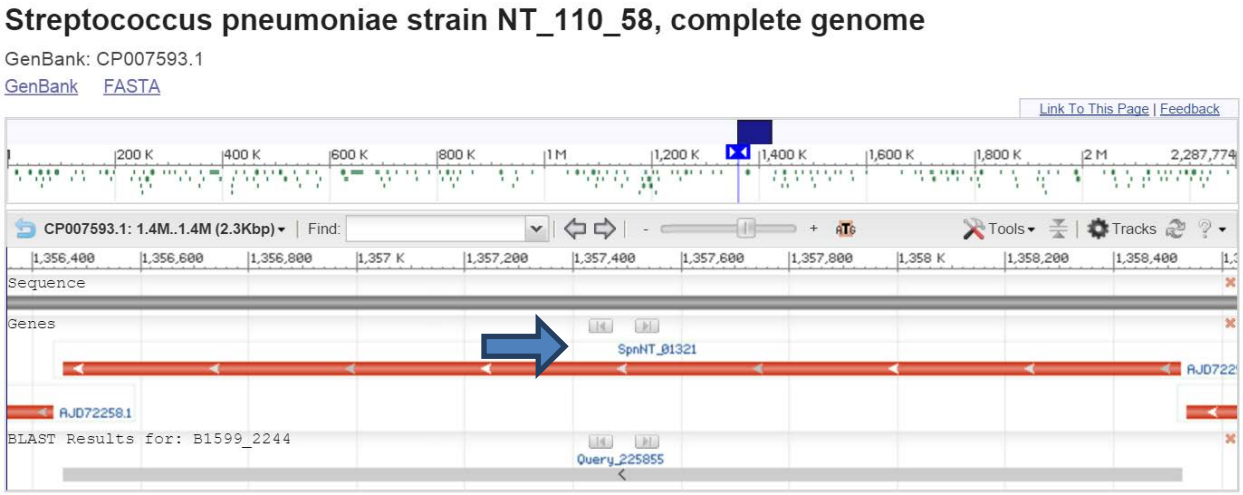
    and three temporary files.

    c. *Positive_nucleotide_sequences.fasta*
    d. *Negative_nucleotide_sequences.fasta*
    e. *RC_Negative_nucleotide_sequences.fasta*

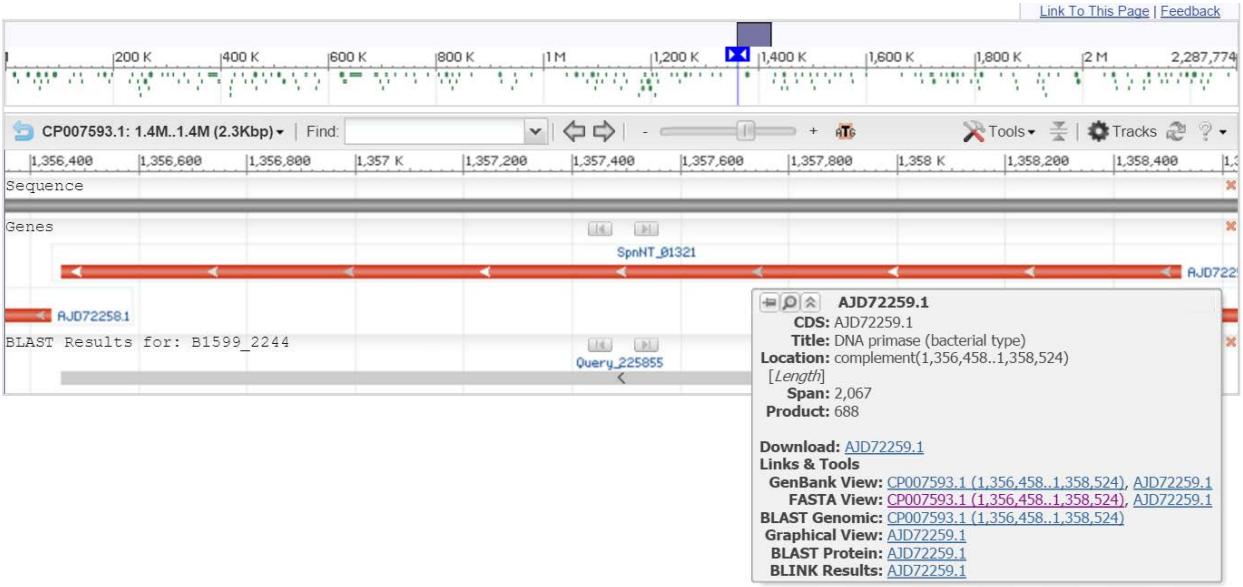
You will generate these sequence files using the text editor of your choice.



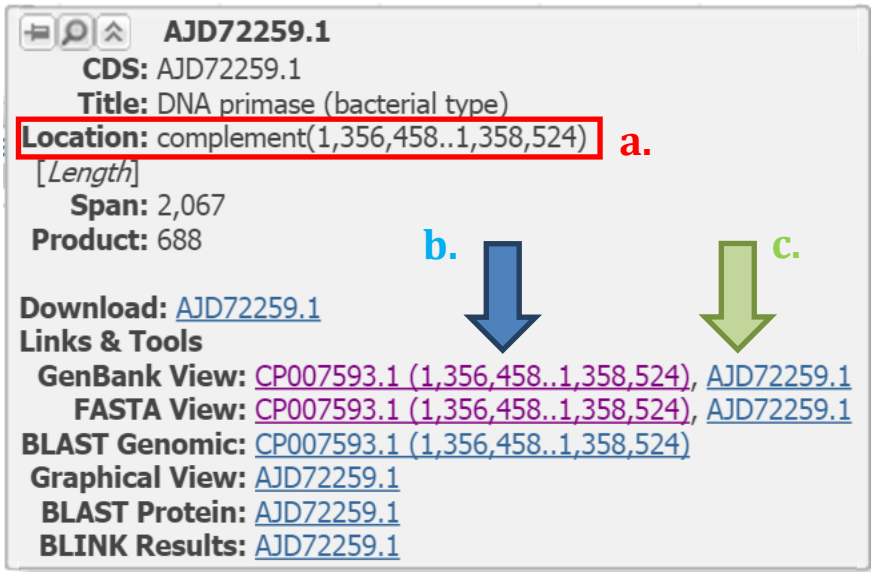
2. For each sequence on your list, go to the Alignment entry and click the 'Graphics' Link (red arrow, above figure).



3. This will open the graphical browser with the sequence of interest centered in the window and shown in red. Note the query graphic underneath. Note, the text above the red sequence of interest is the locus tag (blue arrow). The locus tag is also accessible from the Genbank file, as described later.



4. If you hover your mouse over the red sequence graphic an information box will pop up. This box contains three things that are useful:



a. Location Field (red box) – If the location field shows the word 'complement' then the gene is encoded on the 'Crick' or 'Negative' strand and is read right to left. If the word 'complement' is not shown the gene is encoded on the 'Watson' or 'Positive' strand and is read left to right. If you have a gene located on the Crick/Negative strand you will need to perform a Reverse Complement operation on it before aligning it!
b. GenBank View nucleotide record (blue arrow) – This links to the page with the nucleotide sequence.

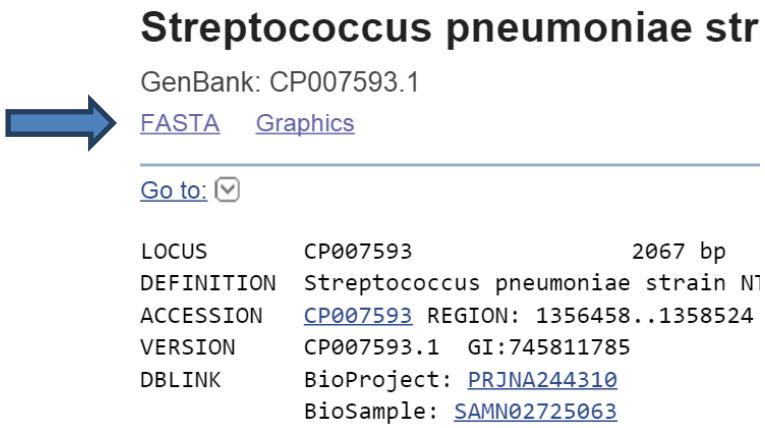
c. GenBank View amino acid record (green arrow) – This links to the page with the amino acid sequence.

5. Copy the locus tag at the beginning of a new line in *Amino_acid_sequences.fasta.* This will be the name for that sequence, so be sure to put a '>' at the beginning of the line; e.g.,

   >SpnNT_01321

   The sequence label must also be added to your nucleic acid sequence file. Copy it at the beginning of a new line in *Negative_nucleotide_sequence.fasta*, if the nucleic acid sequence is a 'complementary' sequence, or at the beginning of a new line in *Positive_nucleotide_sequence.fasta,* if it is not.

6. Click on the link to the GenBank View Nucleotide record.



7. The Genbank record will open. Click the FASTA link to see the sequence in FASTA format. Copy the sequence into *Negative_nucleotide_sequence.fasta*, if it is 'complementary', or into *Positive_nucleotide_sequence.fasta*, if it is not.
8. Return to the page with the graphical genome browser and open the GenBank View amino acid record. Click the FASTA link, and copy the sequence into A*mino_acid_sequence.fasta*.

Repeat this process for each sequence on your list. When complete, save your *sequence.fasta* files.

OBTAINING THE REVERSE COMPLEMENT OF A SEQUENCE

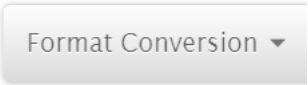
DNA is a double-stranded molecule. In most genomes, coding sequences are found on both strands. The preponderance of genes are encoded on one strand, sometimes called the positive or "Watson" strand, but genes are also found on the negative or "Crick" strand. For example, the red arrow in this DNA fragment represents an ORF on the Crick strand.

```
TAC ATG ATC ATT TCA TGG AAT TTC TAG CAT GAA
ATG TAC TAG AAT AGT ACC TTA AAG ATC GTA CTT
```

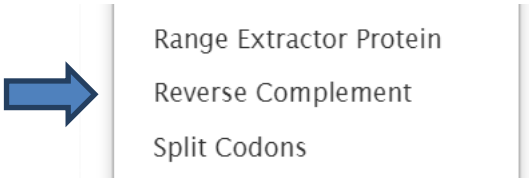
Database sequences always go in the same direction, which means that if you download a gene on the negative strand, it will be upside down and backwards. In the example above, the sequence you would obtain from the FASTA page would be `ATTTCA...CAT`, *not* `ATGCTA...TAA`. If you have such sequences in your data set, you will have to flip them over and invert them using the following steps:

1. Open this link: https://www.dna20.com/resources/bioinformatics-tools

    a. Click the Format Conversion Menu option near the bottom of the screen.

    b. Click the Reverse Complement option from the menu.

    c. Copy and paste your sequences from your *Negative_nucleotide_sequences.fasta* file into the text box, check that the drop-down box reads 'reverse-complement', and then click the Submit button.
    d. A new window will pop up that contains the reverse-complement of your original sequences. Copy and paste these into your *RC_Negative_nucleotide_sequences.fasta* file and save this file.
    e. Finally, copy and paste the contents of both the *Positive_nucleotide_sequences.fasta* and *RC_Negative_nucleotide_sequences.fasta* files into *All_nucleotide_sequences.fasta* and then save this file.

GENERATE AMINO ACID AWARE NUCLEOTIDE ALIGNMENTS.

PAL2NAL is a way to align nucleotides in a way that takes into account the amino acid translated from each nucleotide codon.

1. Use your preferred method to generate a multiple sequence alignment of your amino acid sequences. Save it in a file called *amino_acid_alignment.aln* for posterity.
2. PAL2NAL requires that each entry in the amino acid data set is in the same order as the entries in the nucleotide data set. ***This step is critical.*** By clicking the Sequence Name in Jalview and using the Up and Down Arrow keys you can re-arrange the order of your sequences to match the order found in your *All_nucleotide_sequences.fasta* file.

   Alternatively, you can read your *All_nucleotide_sequences.fasta* into Jalview. Sort the amino acid sequences and the nucleic acid sequences by their labels (Calculate->Sort->By ID). Now they will be in the same order.

3. Once your MSA is generated and in the correct order, go to 'File > Save As…' and name your MSA and then set the file format to **Clustal (.aln).**
4. Open the PAL2NAL page: http://www.bork.embl.de/pal2nal/
   a. Scroll down to Input File 1, this is for your Amino acid alignment. Click the 'Choose File' button to upload the Clustal file you made in the previous step.
   b. Scroll down to Input File 2, this is for your Nucleotide sequences. Click the 'Choose File' button to upload your *All_nucleotide_sequences.fasta* file.
   c. Scroll down to the Option Setting section. Under the Codon Table you will want to choose: "Bacterial, archaeal and plant plastid code (NCBI: transl_table=11)"
5. You can now click the 'Submit' button, the window will refresh and the output will be displayed. If PAL2NAL completed without error you can copy your nucleotide alignment by copying everything after the line "CLUSTAL W multiple sequence alignment".
6. Open a file in your text editor called '*Complete_nucleotide_alignment.aln*', copy and paste the PAL2NAL output into it, and save it.

You are now ready to trim your alignment and perform model selection and tree building.