03-727 Project Description

Milestone 1

Your *first milestone* is to determine the phylogenetic distribution of your gene within the Streptococcus genus and in more distantly related bacteria; and select an appropriate set of sequences for phylogenetic analysis.

We will devote Thursday's class (Nov 19) to these tasks. You should come to class prepared with the following:

- Bring your laptop to class, with the blast output from each of your searches open in a browser. To prepare for this, you should conduct at least one search of the nucleic acid data base and at least one search of an amino acid data base, using your assigned sequence as a query. Based on the results of the BLAST searches, you should also have:
- A proposed list of sequences for a DNA-based phylogenetic analysis.
- A proposed list of amino acid sequences for a protein-based phylogenetic analysis.

These items are a deliverable for your final project and will be considered in determining your project grade.

Some guidelines:

<u>Query sequences:</u>

You can download your sequence here:

http://www.cs.cmu.edu/~durand/Phylogenetics/2015/Projects/Sequences/<yourFirstName>.fasta

For example, Pieter's sequence is in the file *Pieter.fasta*.

<u>What should you be looking for?</u>

1. Recall that our primary phylogenetic question is "Did horizontal transfer contribute to colonization of the human eye by Streptococcus pneumoniae?" Your sequences were chosen because they are found in most sequenced Streptococcus pneumoniae genomes that were sampled from conjunctivitis patients and are absent from other Streptococcus pneumoniae genomes. One goal of this project is to determine where they came from.

In order to find genomes that may be the origin a horizontally transferred gene you should focus on DNA sequences that are full length matches to your query. For very closely related sequences, these will be obvious. If your family is very diverged, you may also find sequences that cover most of the query sequence, but with several interrupted alignments, instead of one long, contiguous alignment.. You may also find strong matches to paralogous sequences, which are interesting for other questions, but are not likely to be the origin of a horizontally transferred gene.

2. A second goal is to use phylogenetic analysis to characterize the gene family of your sequence. In this case, both orthologs and paralogs are of interest. Depending on the divergence of your family, you may find DNA or amino acid searches work better to collect a data set for this second goal.

Strategies for searching

General:

- Experiment with modifying the parameters of your search to see what values give you the best sensitivity.
- Recall that the statistical significance of a retrieved sequence depends on the size of the query and the size of the data base. How are these quantities related? How could you modify your search to make this work in your favor?

Nucleic acid searches:

- Use the "blastn" program to search with your nucleic acid sequence.
- Under "Program Selection", I recommend that you use the "Somewhat dissimilar sequences" option.
- You can reduce the number of false negatives by reducing the word size in the "Algorithm parameters" section. This may retrieve sequences that were not returned with the default parameters. It will also result in a slower search.
- You may also be able to reduce the number of false negatives by adjusting the Match/Mismatch score.

Amino acid searches:

- Your sequences are nucleic acids. To carry out a search against a protein database, you have two options.

1) Search with the "blastx" program, which searches protein databases using a translated nucleotide query. Just paste your sequence into the "Query Sequence" box.
2) Translate your sequence to obtain an amino acid sequence and then search with the blastp program you used in the practicum. Various tools for translating cDNA sequences are available. For example, Jalview has a "Translate cDNA" function under the "Calculate" menu.

- You may want to consider searching the "refseq_protein" database, instead of nr. (Why?)

- Again, you can reduce the number of false negatives by reducing the word size in the "Algorithm parameters" section. This may retrieve sequences that were not returned with the default parameters. It will also result in a slower search.