The find-a-gene project

Modified from the original version of Prof. Pevsner

The find-a-gene project is a required part of this course. You should prepare a written report in a Word document that has the following components. The assignment is due the last day of BIFX550. Steps [1] to [4] can be accomplished very quickly (at best within 5 or 10 minutes), so if you don't succeed at first, just keep trying. The main point is for you to grasp the principles of database searching and sequence analysis that we cover in course.

[1] Tell me the name of a protein you are interested in. Include the species and the accession number. If you do not have a favorite protein, select a protein that is associated with a disease.

[2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include the output of that BLAST search in your document. If appropriate, change the font to Courier size 10 so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC). It is not necessary to print out all of the blast results if there are many pages.

On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score.

In general, step [2] is the most difficult for students because it requires you to have a "feel" for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e. a sequence that is not "novel"), a near match (something that might be "novel", depending on the results of step [4]), and a non-homologous result.

[3] Gather information about this "novel" protein. At a minimum, show me the protein sequence of the "novel" protein as displayed in your BLAST results from step [2]. In some cases, you will be able to do further BLAST searches to obtain even more sequence of your novel gene.

Here, tell me the name of the novel protein, and the species from which it derives. It is very unlikely (but still definitely possible) that you will find a novel gene from an organism such as *S. cerevisiae*, human or mouse, because those genomes have already been thoroughly annotated. It is more likely that you will discover a new gene in a genome that is currently being sequenced, such as bacteria or primates or protozoa.

[4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, "novel" is defined as follows. Take the protein sequence (step [3]), and use it as a query in a blastp search of the nr database at NCBI.

--If there is a match with 100% amino acid identity to a protein in the database, from the same species, then your protein is NOT novel (even if the match is to a protein with a name such as "unknown"). Someone has already found and annotated this sequence, and assigned it an accession number.

--If there is a match with less than 100% identity, then it is likely that your protein is novel, and you have succeeded.

--If there is a match with 100% identity, but to a different species than the one you started with, then you have succeeded in finding a novel gene.

--If there are no database matches to the original query from step [1], this indicates that you have partially succeeded: yes, you may have found a new gene, but no, it is not actually homologous to the original query. You should probably start over.

[5] Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of other members of this family. A typical number of proteins to use in a multiple sequence alignment is a minimum of 5 or 10 and a maximum 30, although the exact number is up to you.

[6] Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use any program such as MEGA, PAUP, or Phylip.

MEGA software: https://www.megasoftware.net/

Jalview software: http://www.jalview.org/

[7] Compare the predicted structure of your protein to that of a known structure. You can use the online protein prediction servers for modeling.

Phyre2: http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index

Very easy to use. After you have discovered your novel protein, you can take the fasta sequence and paste it into the box from the server to get a model (3D)

[8] Optional: show whether this gene is under positive or negative evolutionary selection.

[9] Discuss the significance of your novel gene. What have you learned about this gene/protein family?