Lab 16: Comparative Genomics

Introduction

There are many resources for exploring genes, gene families and genomes on the web. Some use a phylogenetic approach to aid in the analysis of gene family evolution, though many do not. There is not enough time to do an exhaustive search of phylogenetically based genome analyses on the web, let alone of all the available web sites for genome analysis in general. I would recommend investigating these and other resources further on your own time, if you are at all interested in the subject.

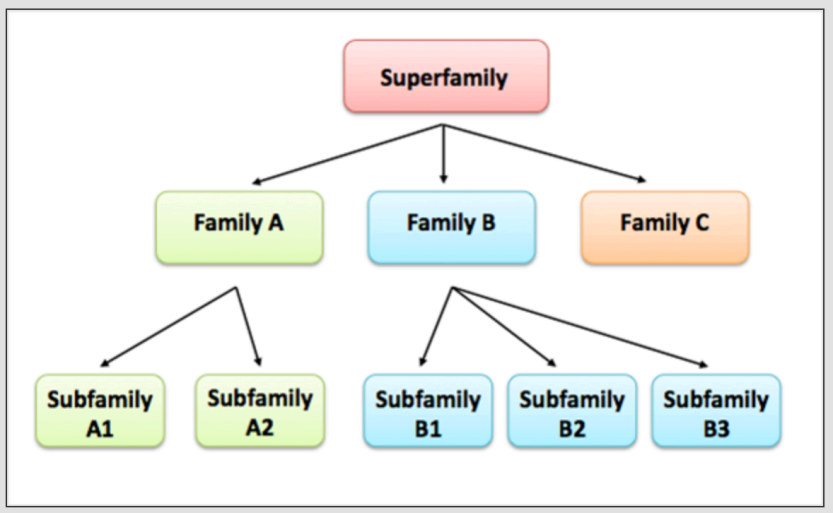
Today we'll discuss some bioinformatics portals and explore some of the tools available here. The goals of the lab today are to:

I. Introduce you to Bioinformatics Portals, specifically ExPASy
II. Analyze a known amino acid sequence
III. Deduce information regarding a mystery amino acid sequence

Exercise 1: Online Bioinformatics Portals

Classifying proteins into families and identifying important domains and sites is invaluable for helping biologists to identify distantly related proteins and to predict their functions. A daunting array of resources, each with different strengths and weaknesses, is now available for searching genomes and proteomes with 'protein signatures' – diagnostic entities that are used to recognize a particular domain or family.

What is a protein family? A protein family is a group of proteins that share a common evolutionary origin reflected by their related functions and similarities in sequence or structure.



Protein families are often arranged into hierarchies, with proteins that share a common ancestor subdivided into smaller, more closely related groups. The terms superfamily (describing a large

group of distantly related proteins) and subfamily (describing a small group of closely related proteins) are sometimes used in this context.

There are two main Bioinformatics portals that I would like you to look at briefly today. We will primarily work with some tools through ExPASy today, but if there is a particular question you are interested in, EMBL-EBI might have some additional tools for you, so I'm listing it here.

ExPASy

http://www.expasy.org/

EMBL-EBI

www.ebi.ac.uk/services.

Question #1: **What do each of these acronyms stand for? What type of tools do these portals offer? Take a few minutes to explore each of them.**

Exercise 2: Examining Protein Function Using UniProt

Within both of these portals you will notice some familiar friends – Clustal, BLAST, etc. If you explore these two portals, you will also notice that there is a good amount of overlap between the tools. We'll start today using ExPASy and some of the tools offered here. From the homepage, click on the program **UniProtKB** found under the "Popular Resources" column on the right hand side of the page.

UniProt is a resource to examine protein sequence and function information. Click on the "Blast" tab at the top of the page.

Download the file Nematode lin-39.fasta from the IB200 website. Open it in a text editor and you'll see a Nematode AA sequence for a Hox gene. Copy this and paste it into the box at the top of the page under the "Blast" tab. Click the Blast button – this will take a minute or so to run.

Once finished, UniProt will display all information regarding the functional information related to this protein. Scan down the page and click on the entry that has 100% Identity to your sequence.

Question #2: **What is the name of the protein that you searched? What is the function of this protein? Explore this page and figure out a way to see a 3-D model of this protein. Take a screenshot of this 3-D model and send it to me.**

You will also find a section regarding Gene Ontology (GO). Gene Ontology is a controlled vocabulary used to describe the biology of a gene product in any organism. There are three independent sets of vocabularies, or ontologies, that describe:

1. the biological process in which the gene product participates
2. the cellular component where the gene product can be found
3. the molecular function of a gene product

You may also be interested in how your protein might interact with other proteins. Scroll down to the "Protein-protein interaction databases" and select **STRING**. STRING is another useful tool and you will see an interaction network at the top of the page. Click on the central bubble corresponding to your protein of interest, lin-39. You'll see more information about your protein and you can now also see what it looks like (in case you had trouble earlier finding this information). If you scroll down the page, you will see a list of Predicted Functional Partners. Using this information and the bubble networks towards the top of the page, answer the following:

Question #3: **Given that you ideally want independent data sets and characters to reconstruct phylogenies, what thoughts do you have regarding the functional partners shown here?**

Okay, now go back to the UniProt page. Now we'll explore a Phylogenomic database, which use a phylogenetic approach to analyze gene family evolution. We'll look at the tool labeled **HOGENOM**.

Question #4: **What is the name of the gene family? What alignment program was used to build this tree? What optimality criterion was used to build this tree?**

Exercise 3: Deducing the Function and Gene Family for a Mystery Sequence

Download the file Mystery.fasta from the IB200 website. Open it in a text editor and you'll see a Mystery AA sequence. Run this sequence through **UniProt** and select the taxon with the entry that has 100% Identity to your sequence. When examining Phylogenomic databases, HOGENOM is not listed, so instead use **HOVERGEN**. Then use this to answer the following questions.

Question #5: **What species does this sequence belong to? What is the function of this protein? What alignment program was used to build the tree in HOVERGEN? What optimality criterion was used to build this tree? What species was found to be sister to your sequence of interest? Figure out a way to download this tree, read it into FigTree, display bootstrap supports, and highlight the queried taxon in the tree.**