Lab 9: Web Applications for Gene Family Evolution

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, many do not. Today we will investigate just two pages that do. 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.

To turn in for today: nothing. Your priorities should be to (1) successfully complete the BEAST tutorial from the dating lab, and (2) finish the molecular evolution lab which used RAxML and MrBayes.

Part 1: GenBank and online BLAST, and Jalview

NCBI's GenBank has an immense amount of data available – gene sequences, and lots of other stuff. These data can be accessed through web forms, FTP, command-line programs (google "BLASTtools", you can search a local sequence database or a remote online one), BioPython (an extension of Python for bioinformatics), BioPerl (a similar extension, but Python is better), and (I think) BioConductor (a huge R package designed by bioinformatics people, mostly for expression data though, apparently not genomics/phylogenetics).

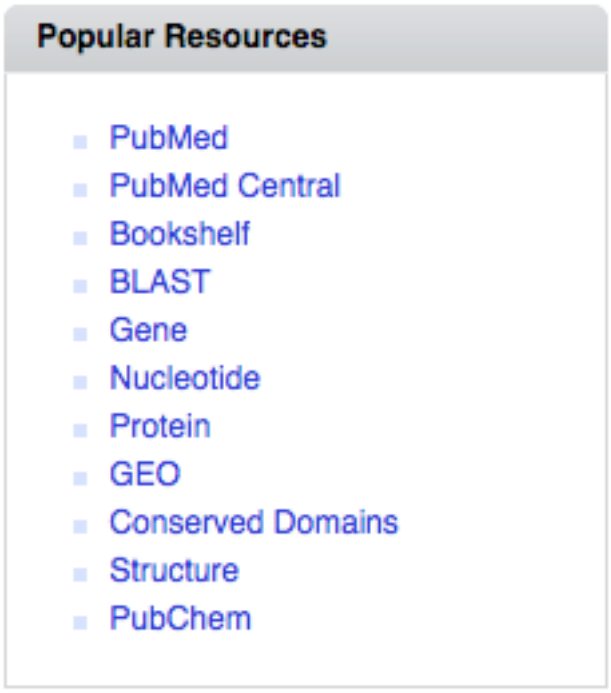
BLAST is an extremely rapid search tool used to find sequences related to a target sequence. It is basically Google for sequence data, although it has many more formal uses. Everyone should have basic BLAST abilities in their skill set, so I am including a chunk of the GenBank/BLAST lab from a IB200A lab. *Note: Please skip this section, if you did the BLAST/Jalview exercise in IB200A last year, or consider yourself thoroughly familiar with online BLAST and viewing alignments.*

Questions: figure out the answers to the questions, but you don't have to turn anything in.

GenBank

First, we'll try out some of GenBank's functions. Open your web browser on your computer and go to the site http://www.ncbi.nlm.nih.gov/. This is the National Center for Biotechnology Information website, where GenBank (among many other things) resides.

It is important for you to be aware of the diversity of resources available at NCBI. Under "Popular Resources," you will see the following:



Question 1: Click on each of these, figure out what they basically do, and write a (short!) description of each above.

Note that there are many, many other resources!

Nucleotide database

Select 'Nucleotide', GenBank's DNA/RNA sequence database. Try typing in the name of your favorite taxon in the search bar to see if there are sequences for it. Once you've done this, a list of sequences will appear (if there are sequences for your organism).

- You can also search for by taxon using the 'Taxonomy' database. Type the name of e.g. a family.
- To see all of the nucleotide sequences available for your taxon, check the "Nucleotide" box, then "Display."

Question 2: What does NCBI say about its taxonomy data at the bottom of your list of found taxa?

- Click on the number next to your taxon of interest. This displays the matching sequences. Each sequence is listed by its accession number, and information about the taxon, gene, etc. is also provided. Follow the link for one of the sequences you've found. A new page with various information about the authors of the sequence, the taxon, gene, where it was published, etc. will appear. At the bottom of the page you will find the sequence itself. Near the top of the screen, you can see that there are several options for displaying and saving the sequence. Check out some of the display options (choose them from the pull-down menu and then push display), but don't bother saving anything for now. If you're looking for sequences by a particular author or a particular gene, you can

also type in those or any combination of them and do a search. Feel free to try this if you like.

- Pick a sequence that you think would be a good one use in a phylogenetic analysis of your group (e.g., a sequence that looks like it has been sequenced in many of the relevant species, that is conserved, named, etc.).
- Figure out how to save it to a FASTA file on your hard drive.
- Also, just cut-and-paste the sequence from the webpage to a text file.

BLAST

Now we'll try a BLAST search on the sequence you just found. BLAST searches are useful for finding sequences similar to one you have generated or found.

Question 3: What does BLAST stand for? What decision are you making by searching for sequences with the BLAST algorithm instead of some other algorithm (hint: consider what the 'LA' means, and what other options there might be among search algorithms.

- Open the BLAST homepage **in a new window** (http://blast.ncbi.nlm.nih.gov/Blast.cgi), and then click on the 'nucleotide blast.' This is the option for searching for nucleotide sequences with a nucleotide sequence, but other options (such as searching for translated sequences, searching within the human genome, or searching for really close matches quickly) are available.
- Now copy the sequence you found in GenBank, go back to the BLAST site and paste it into the 'search' box. (Hint: the numbers get ignored, just the letters are read.)
- Pick an appropriate database to search.

Question 4: What's the default database? What database did you decide was appropriate to search?

- When you've done that push the BLAST button. The search may take a couple of minutes, so be patient.
- Once the search is done, you can check out which sequences were found that generated significant alignments with your query sequence by scrolling down the page. You can also see the alignments with these sequences that the BLAST algorithm generated as well. There is a graphical representation (near the top of the results page) that shows where the various hits could be aligned with the query sequence and how good that alignment is.

  Question 5: How many hits did you get? Did the taxa that "should" have been the closest phylogenetic relatives, based on taxonomy, all come up as the closest matches to your sequence? If not, what are some possible reasons why not?

  Question 6:
  - What does "e-value" stand for? (look it up online if necessary)
  - What does that value mean?
  - What is a good e-value, and what is a bad e-value?

- Does an e-value represent Manhattan distance or Euclidean distance between two sequences?
- How should you modify an e-value if you conduct 100 searches instead of 1? How about 1000 searches?

- Finally, click on the 'Taxonomy Reports' link if you want to learn more about the organisms that matched the query sequence. Obviously, you can do a lot more on GenBank. Feel free to explore the site further if time allows.

Jalview

You can download Jalview at: http://www.jalview.org/download.html

- Now we'll search for and download the sequences that we'll use in Jalview. Either use a manageable number of sequences from your own group (say, 5 to 100)
- Or do the following (these turtle sequences will be used as the example for the rest of the lab, but you can modify accordingly): Go back to the main GenBank web page, and search in 'Nucleotide' for "*emydidae feldman*" this is the taxon (turtles) and the author (Feldman, who submitted the sequences.) When the results appear, select the cytochrome b gene for *Terrapene carolina* (the accession number should be AF258871), *Emydoidea blandingii*, *Chrysemys picta, Clemmys guttata*, and *Clemmys marmorata*, pick 'FASTA' from the display menu and then 'file' from the send to menu. Save the file to your desktop. Name it yourname_turtles.fasta.

Now that we have our sequences, we can do some aligning. The techniques we will be using in Jalview are relatively simple. The program has numerous other functions that we will not use today, but that are useful for exploring various properties of molecular data. If you are planning on including molecular data in your project, you may wish to explore these options further by using the extensive Help information included with the program.

- Open Jalview from the desktop. After the program starts, three windows will appear. Close them all. Now go to the file menu and select 'Input Alignment' > 'from File'. A dialogue box will appear. Change the Format to 'All Files'. Select your saved sequences (yourname_turtle.fasta) and click 'Open'.
- Once you've imported all of your sequences, they will appear in the alignment window and a consensus sequence will appear along the bottom. Each sequence will be identified by its accession number.
- GeneDoc can shade the nucleotides in several different ways, showing different properties of the sequences. Pull down the color menu and select 'Percent Identity' which indicates what percentage of the residues in a column match the consensus sequence. Columns that are shaded dark blue are more than 80% conserved, columns that are blue are more than 60% conserved, columns that are light blue are more than 40% conserved, and columns that are white are less than 40% conserved. As you can see, even without doing any additional aligning, these sequences have large conserved regions, which is not surprising given that these turtles are relatively closely related.

Many of the other shading options have to do with what types of Amino Acids the sequence would code for in a protein sequence alignment. You can translate a sequence using this program, but we won't get into that now. The most commonly used coloring for nucleotides is 'nucleotide'. This colors the sequence according to the nucleotide identity.

- Now switch to 'Percent Identity' and scroll down through the different blocks of sequences. As you can see, the sequences generally match up pretty well for most of their length, except at the end where the *Chrysemys picta* sequence is notably different than the others. Also note that this sequence is four base pairs longer than the others. We could simply leave the sequences as they are, but we might be able to do some additional aligning to get us closer to the true phylogenetic signal.
- One thing we might do is use ClustalW through Jalview to align all of our sequences automatically. Go to the 'Web Service' menu and select 'ClustalW Multiple Sequence Alignment.' This will align all the sequences using ClustalW online, which we'll deal with more later. Once it is done, be sure to check out the area near the end of the sequence, which is where most of the changes took place. As you can see, the new alignment added a few gaps, but resulted in a much closer fit between this sequence and the others.
- When you are finished with your alignments, you may wish to save your work to import it into other programs (*e.g.*, PAUP*, MrBayes, etc.). Go to the 'file' menu, and select the 'save as' option. You can see that several formats for saving are available. Choose the .fasta format and name your file aligned_turtles.fasta.

Part 2: PhyloFacts

PhyloFacts is an on line structural phylogenomic encyclopedia maintained by the Berkeley Phylogenomics Group. Its web site can be found at http://phylogenomics.berkeley.edu/phylofacts/. It works by doing a simple search for sequence similarity between a given sequence and a number of "books", which are groups of related sequences. It then reconstructs a gene tree in order to identify the gene family and subfamily. The phylogeny is used to compare to other data bases and identify functional domains and protein structures.

Let's check out some of the protein families that they've already analyzed. Libraries are sets of sequences, either organized around related genes, gene function or taxa. Click on **"GPCRs"**. These are G-Protein Coupled receptors, a clade of common eukaryotic trans-membrane proteins that receive a diverse set of extra cellular signals and in response stimulate an intracellular signal through the activation of membrane bound G-proteins. On this page you will see two different options. "Protein Search" allows you to classify your own protein sequence; we'll do that in a minute.

Let's take a look at some of their "Books". Books are not long texts on gene families or phylogenomics. Instead they are groups of related genes with a great deal of associated information. Start by looking at a small one. How about "Cannabinoid Receptors", there are only 16 of them in their library. Cannabinoid receptors are the most common type GPCRs found

in the brain, but they are also expressed throughout the body including in the lungs, liver, kidneys, immune system and male gametes.

The first thing you will see is an accession number and name for the gene family. Below that you will find a figure describing the domains found in the gene. The legend explains that the little red boxes are transmembrane alpha-helices and the blue line is the seven transmembrane domain (7TM). Not surprisingly the 7TM covers the 7 alpha-helix domains.

We'll skip all those boxes with interesting options and move on to the summary of the genes. This includes the number of sequences and the length of the genes. Below that is the taxonomic distribution for these genes. Still further down you will find the GO descriptions of the gene under three different sets of categories. Clicking the sequence annotations link will provide you with the same information for each gene in the alignment. Finally there is a reference for a recommended literature review about this protein family.

Now we'll move to those interesting boxes. Click on the box with the tree. This is a Neighbor Joining tree. (Ugh, not NJ!) This tree looks OK. There are two major clades of Cannabinoid Receptor genes. It looks like both genes have expanded in the human lineage. The relationship of these genes within each clade pretty much matches the relationship of the taxa. I don't know about the cat being sister to the rodents. If there are any taxa that you don't know click on the taxon name to get a description. For example, what is *Taricha granulosa?* Oh, a rough skinned newt. Well that should probably be sister to *Xenopus*, but what do you expect from a NJ tree? You can also see that information by checking the box next to "Common Names".

Some of these genes do seam very out of place. Why is the one puffer fish gene all the way at the base of the vertebrates? This implies that there was a duplication in the common ancestor of the teleosts and the tetrapods, and that one of those genes was lost independently in the zebrafish and the tetrapods, but retained in the puffer fish. Possible, but unlikely. My guess is long branch attraction. Also, there are only primate and rodent genes in the second clade of Cannabinoid receptors. This could be cause these genes were lost in the other taxa independently, but I would favor one of two other explanations. One possibility is that this was a duplication in the common ancestor of rodents and primates, but that this tree sucks and this clade of genes should really be nested inside the other. A more likely possibility is that the orthologs of these genes have not yet been identified in the other taxa. "Absence of evidence is not evidence of absence", especially if you haven't looked really hard. Finally there is that one zebrafish gene outside the rest of the genes; I don't know what's up with that one. Is it just there as an outgroup?

So, let's look at a better tree. Close down that window, so that you are back to the general Cannabinoid receptor window. Pull down the menu below the tree box to "View full ML tree". A new window will appear; expand the window to see the whole tree. This tree does not have all the accouterments of the other tree, but it is probably a better tree. You can see that this tree has a more reasonable set of relationships. The cat gene isn't sister to the rodents, and that puffer fish gene is up there with the other fish genes, even if it's not sister to the other puffer fish gene. The newt gene still appears to be in the wrong place, and I don't know what's up with that zebrafish outgroup. OK, you can close this window.

That whole "Conservation analysis" thing has not been set up for this gene. To see the alignment pull down the menu below alignment to "Quick view-HTML". You can see the protein alignment used to deduce this tree.

Back up to the PhyloFacts homepage; we'll check out a new gene. Click on the "Innate Immunity" link. This is a group of genes classified by function, not relationship. All these genes are involved in innate immunity, meaning immunity that does not require previous exposure to the pathogens. These genes come from taxa throughout the tree of life, and although many of these genes are related, the library comes from diverse classes of proteins. Let's check out a book that's big, but not too big. How about "ABC transporter I"? ATP-binding cassette transporters (ABC-transporters) are an ancient clade of genes found in all three kingdoms that transport all sorts of different things across membranes utilizing energy from the hydrolysis of ATP. This book only represents one subclade of ABC-transporters, which is involved in innate immunity.

First, let's look at the plot of domains. There are three different classes of domains described in this plot. There are a bunch of transmembrane alpha-helices and several of these make up the ABC-transporter trans-membrane domains What is the "ABC tran"? Click on that link. This links to the Sanger Trust Institute database on protein families. This page is specifically for this domain. So, it's the transporter part of the protein and it is also involved in ATP-hydrolysis. Click on the "Domain Organization" tab over on the left. This shows a list of all the other domains that this domain is found in association with and the different combinations of those domains. Can you find any that are organized like our protein? Click on the "HMM Logo" button. This plots the protein sequence; larger letters represent residues more commonly found at those sites. You could try checking out some of the other options, if you want. When you're done, shut down this window and go back to the ABC-transporter page for PhyloFacts.

Click on the tree icon on the ABC-transporter page for PhyloFacts to view an NJ tree. The ML tree is better, but the NJ tree is prettier. As you can see, this is a large tree covering many different taxa and many paralogs from several of those taxa. It's almost too big to really take in. That's why I had you look at the other tree first. Pick out some clade where you have some knowledge of the taxon phylogeny. Does the gene phylogeny makes sense? What combination of duplications and losses would be required to make that part of the tree work? OK, shut that window down.

Click the "View Structures" link under the 3D picture of the protein. Oooh, see the cool picture. Let's skip that for a minute and go down to the bottom of the page. You'll see a plot of the significance scores against position in the protein sequence. The x-axis is the sequence with the most common amino acid at every site listed. The higher the significance score, the more commonly that amino acid is found at that site in this protein family. If an amino acid is more highly conserved, it implies that it is being maintained by natural selection and thus has an important function.

Now we can go up to the top of the page and look at the pretty picture. That is a 3D picture of our protein. Right click on the picture select **spin** and turn it **on**. Bring up that menu again select **zoom** and **800%**. Cool, now we can see what's really going on. Do you see the alpha-helices and beta-sheets? The purple residues are sub-family significant and the yellow residues are significant at both the family and sub-family level. I would call this a "semiphylogenetic" method for identifying important functional AAs. It is "phylogenetic" in that it does consider relationships like families and subfamilies, but it is only "semi-" in that it does not consider the way these characters change on the tree. Try out some other viewing options by right clicking on the image. They don't all work. I would recommend **Render>Scheme>Ball and Stick**. I don't know what that other big structure is. Maybe a protein that interacts with it.

OK, shut all that down and go back to the PhyloFacts home page. We're going to try searching with one of our own genes. Click on the "Sequence Search" link. Download the file *Nematode_lin-39.fasta* from the 200B web site; you can find it under the lab prep for this lab. Open it in a text editor; you'll see a Nematode AA sequence for a Hox gene. Copy the whole thing and paste it into the box, and hit "submit". Wait a second for the search to finish. OK, the first thing you will see is a plot showing that PhyloFacts has identified a homeobox domain near the end of the protein. Below that are a list of proteins it has identified with their BLAST scores. The first one is this actual protein. Click on that link. You're pretty familiar with this page. Why don't you navigate around, look at the tree, structure, etc.? Just use this as a tool to learn about this gene.

Question (don't turn it in, but do it): You will find a protein sequence in the file *Mystery.fasta*. What type of protein is this? What domains does it have? What is its function? What clade of taxa is this gene found in? Are there many paralogs in this book, or is it just one or two?