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.

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.fast**a 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.

Phigs

We will look briefly at one other web site http://phigs.org/. Phigs is maintained by the Joint Genome Institute. It also groups genes based on phylogenetic relationship. When constructing their gene trees they try to maintain the appropriate relationships among orthologous genes, such that the relationships among the genes matches the already assumed

relationships among the taxa they are found in. It is not quite as well worked out as PhyloFacts, but it's still pretty cool.

First click on the "Opisthokonts" link. (I don't think that the chordate link works right now.) The first thing you'll see is a list of the taxa from the Opisthokonts with fully sequenced genomes. These are the taxa who's genes the Phigs compare. Click on one of the "Taxa ID" numbers this will send you to the Entrez page for this taxon's (taxis?) genome. This is the NCBI reference for all the information on these genomes. We won't deal with this now, but for real basic analysis it is important to have this reference. Just go back to the previous page.

Down at the bottom of the page there is another little box called "Source 1 Human Details". If you clicked any other taxon's number, you would get the exact same box, but with a different taxon listed at the top. This seams like a particularly poor organization of the data to me. Click on the number "46". Now you will see a list of the taxa involved followed by a table with a bunch of 5 digit numbers defining the rows and smaller numbers defining the columns. The rows are the Phigs and the columns are the taxa. The numbers in the tables are the number of sequences from that taxon that you will find in this Phig. It kinda sucks that the Phig reference numbers don't tell you anything about the Phigs. Click on one of those reference numbers; pick one which has multiple genes in at least a few taxa.

OK, so that first plot you see, I don't really get. It's showing the relationship between every gene, but it's breaking it down to just a couple of types of relationships and I don't see the difference. Below that is a table with information on the different genes in this Phig. Further down are a pair of trees showing the relationships between these genes. What do these trees imply about the history of gene loss and duplication in these lineages?

That's all we're going to do to investigate the different Phigs, go back to the Phig homepage. Click on the "synteny viewer" link. You will see a figure representing Human Chromosome 1. Check the box next to chicken and hit submit. Two different chicken choromosomes will appear on either side of the Human chromosome showing the relationships between the Phigs on these chromosomes. You can clearly see from this plot that Chicken Chromosome 21 and Human Chromosome 1 are homologous over large stretches. There is one large stretch that is in exactly the same order and two others that have completely flipped. Or maybe it's the other way around; ancestral state reconstruction is a bitch with only two taxa. You can view comparisons between other chromosomes in other genomes by changing the source menu, the chromosome number, and the comparison taxa.

Assignment: 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?