Genomic Data Manipulation
BST281 Spring 2019

Species and Gene Trees Activity

Phylogenetic trees provide quantitative, visual representations of the evolutionary relationships between biological sequences (and, by extension, the species from which they derive). Today we will use an online phylogenetics workflow† to build and analyze a few trees.

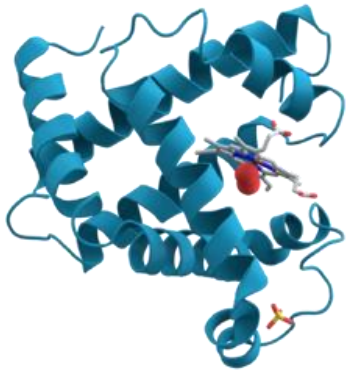
> †Dereeper A.*, Guignon V.*, Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J.F., Guindon S., Lefort V., Lescot M., Claverie J.M., Gascuel O. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W465-9. Epub 2008 Apr 19.

Accessible online at http://phylogeny.fr or http://phylogeny.lirmm.fr/ (mirror).

Part 1: Building a species tree from orthologous myoglobin sequences

Many protein families are conserved across evolutionarily related species, having been passed down from a common ancestor that lived long ago. While the structure and function of such proteins often remain very similar over time, protein primary sequence tends to change (experience substitutions) due to changes in the underlying coding sequence. By comparing modern versions of the protein, we can infer the order in which these changes might have occurred, and by extension the relatedness of their source species.

Today we'll be focusing on myoglobin as a phylogenetic marker:

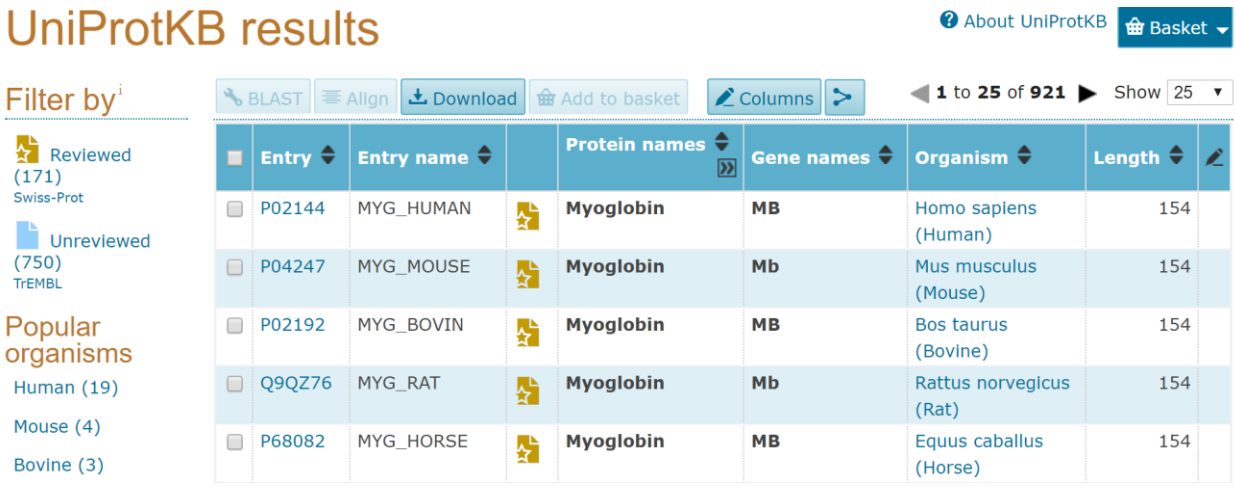


(image from wikipedia.org)

Myoglobin is an iron and oxygen binding molecule found in the muscle tissue of many animal species (including most vertebrates). It also has the honor of being the first protein to have its 3D structure determined (c. 1958), but that's another topic...

We'll start by gathering some diverse sequences from the myoglobin protein family on the UniProt website (http://www.uniprot.org): an excellent resource for protein bioinformatics. Enter "myoglobin" in the search box to find these protein sequences by name.

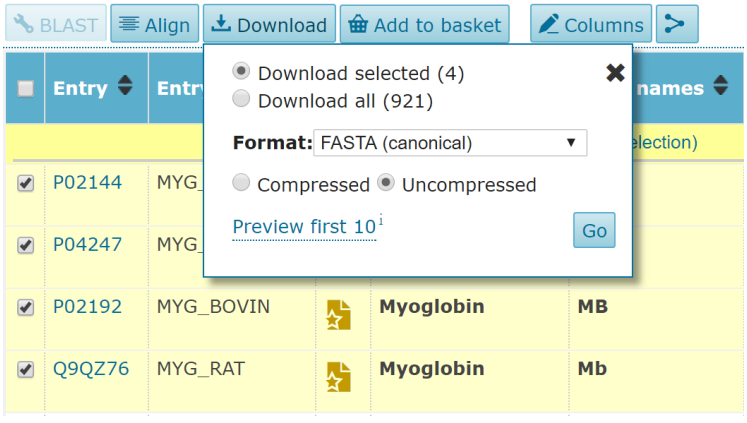
Many sequences are found:



- ***What would be a more principled way to find myoglobin sequences?*** *Assume you have at least one myoglobin sequence (say, the human version) as a reference.*

UniProt divides sequences into two types: *reviewed* (which have benefited from some amount of manual curation) and *unreviewed* (which have been processed by computer only). Clicking on any individual entry (e.g. P02144) will take you to a page summarizing everything that is known about this protein sequence.

Today we are most interested in the sequences themselves. To gather sequences for these proteins, select a handful of them using the check boxes, then click DOWNLOAD and GO (note that the default download format, FASTA, is exactly what we want):



This opens a page of raw text. You can "save as…" this page using the name `myoglobins.fasta`. (Note: my version of this file is available on the course website if you'd

prefer to use that one.) Examine the contents of your `myoglobins.fasta` file. Note that UniProt headers have a very specific format:

```
>sp|P02144|MYG_HUMAN Myoglobin OS=Homo sapiens OX=9606 GN=MB PE=1 SV=2
```

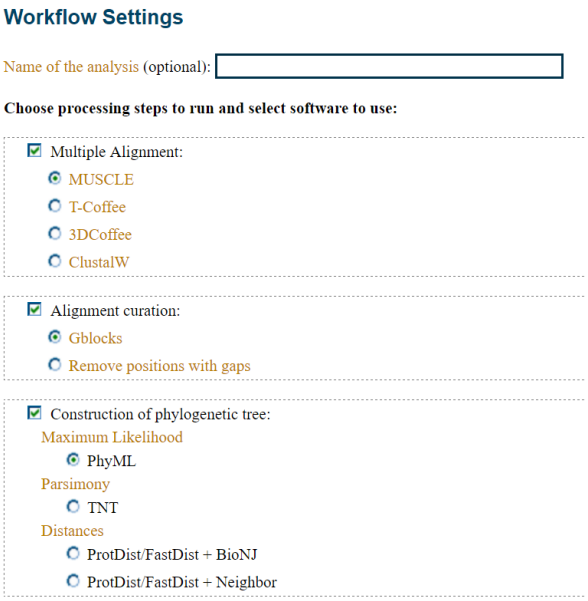
The source organism is revealed in two fields, OS (which gives the organism's scientific name) and OX (which gives the organism's NCBI taxonomic identifier). For tree-building purposes, I recommend replacing the *entire header* with the common name of the source species (using _ in place of spaces). Here, the new header would be:

```
>human
```

You can look up common names for organisms on the UniProt website. Set the search field to "Taxonomy" and enter the NCBI taxonomic identifier:

A separate version of my file that has already been renamed (`myoglobins-renamed.fasta`) is also available on the course website.

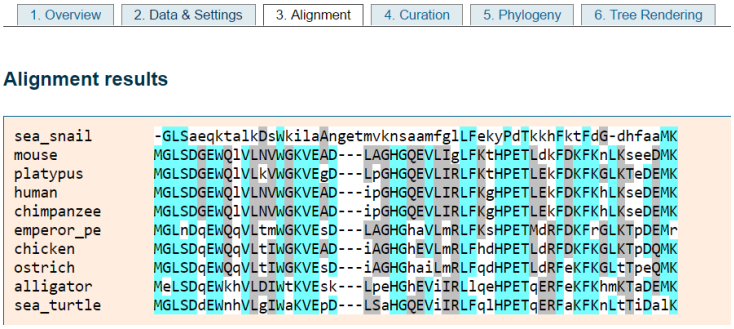
Now we're ready to start building a tree. Point your web browser to the phylogenetics workflow website **(**http://phylogeny.lirmm.fr**)**. Select the appropriately-named "A la carte" option. Note that we now have a variety of workflow "menu" options to choose from when building our phylogeny: 1) choice of methods for multiply aligning our input sequences; 2) choice of methods for curating the resulting alignment; 3) choice of methods for actually building the tree, with three major "families" of methods represented (maximum likelihood, parsimony, and distance-based); and finally 4) choice of methods for drawing the resulting tree.



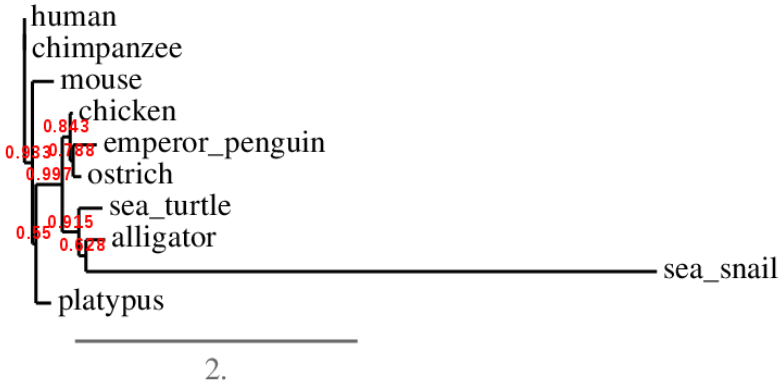
Feel free to stick with the default options for your first tree, which offer a nice balance of speed versus accuracy. Click “CREATE WORKFLOW” to continue.

Upload your `myoglobins-renamed.fasta` file (note that you can also copy and paste FASTA-formatted sequences directly). Scrolling down, you’ll see that you can now fine-tune the individual analysis steps you selected on the previous page. Don’t worry about changing any of these right now, but do note if the tree-building procedure you selected has a “number of bootstraps” option—we’ll come back to this concept later.

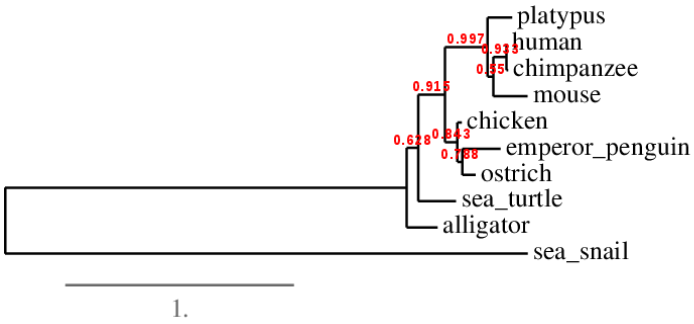
Click “SUBMIT” to launch the workflow, which will take a few minutes to finish (depending on the number of sequences you used and server load). If you’re not too engrossed by the analysis animations, note that the workflow tabs allow you to jump to different points in the analysis to change settings and/or view results. When your workflow moves beyond “Alignment,” click the “Alignment” tab to view the resulting multiple sequence alignment (or MSA):



There are a handful of options here for viewing/coloring the MSA. In this example, bright blue columns reflect amino acid positions that have remained strongly conserved. But we didn’t come here to look at MSAs... Let’s take a look at the finished tree:



While some obvious relationships are evident (e.g. human and chimp appear close to each other), this view is not ideal. Click the button to “reroot using midpoint rooting.” This produces a new view of the tree based on the same underlying data:



Much more clear. Note that the sea snail, the only invertebrate in the group, appears as an *outgroup*. An outgroup is a species that is considerably more diverged (distantly related) from the other species in the tree. The point where the outgroup branches from the tree gives a sense of where "time" begins. (Often times we include a known outgroup on purpose for this reason.) Note that we can explicitly set the sea snail as the outgroup by clicking the "Reroot (outgroup)" button and then clicking on the sea_snail label in the picture (the result is very similar). Feel free to experiment with other style options enabled by the interface.

- **Does the tree match your intuition about the relatedness among these species?**

Red numbers reflect confidence in the accuracy of different branches of the tree, with values closer to 1.0 reflecting greater confidence (the precise meaning of the numbers varies by method).

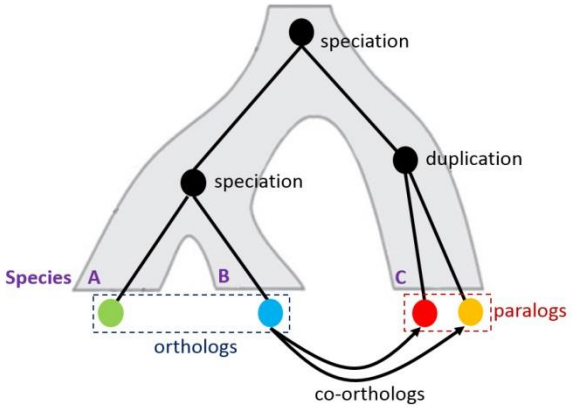
- **Which relationships are the most confident in the tree? Least confident?**
- **Does this tree, inferred from the evolution of a single gene/protein, accurately reflect the evolutionary history of the underlying species? How could we be sure?**
- **Could we include bacterial species in this tree? Explain why or why not.**

If you are interested and have time, try building another workflow with different options and examine the resulting tree. Does it agree with your original tree? Are there some branches that are more/less confident than they were before?

Part 2: Building a tree for the hemoglobin gene family

The previous example focused on orthologous sequences: versions of a gene from multiple modern species that 1) descend from a common ancestor species' version of the gene and 2) which generally perform similar functions across species. Because of these relationships, orthologous sequences are useful for inferring evolutionary relationships among species.

Another class of related gene sequences occurs within a single genome: paralogs. Paralogous genes are related through gene duplication events. In other words, an ancestor of a modern species had a single copy of a gene, which—somewhere in the course of evolution of modern species—was duplicated within a genome to produce another copy. In some cases, paralogous genes become specialized versions of their gene ancestor (e.g. one adapted to colder conditions and another adapted to warmer conditions). In other cases, one of the paralogous genes retains the function of the original gene, and the other adopts a new function.



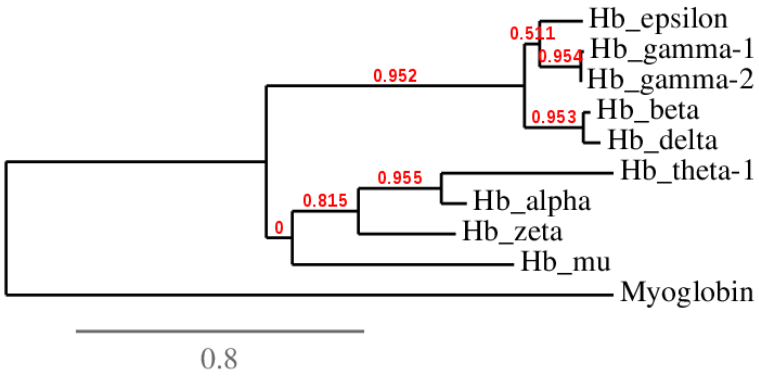
(image from beacon-center.org)

The "globin" family of proteins (of which myoglobin is an example) has experienced this form of expansion over evolutionary time. Indeed, the human genome contains many genes in the globin family that you are probably familiar with: the hemoglobins. While the alpha/beta hemoglobin complex is the best known (occurring in adult blood tissue), other variants of hemoglobin are expressed in non-blood tissues or during embryonic development (naturally, all are encoded in the human genome).

I have gathered a selection of these human hemoglobins from UniProt. The original sequences and their renamed equivalents are available on the course website as `human-globins.fasta` and `human-globins-renamed.fasta`, respectively. Note that I have also included human myoglobin in these files, even though it is not a hemoglobin *per se*.

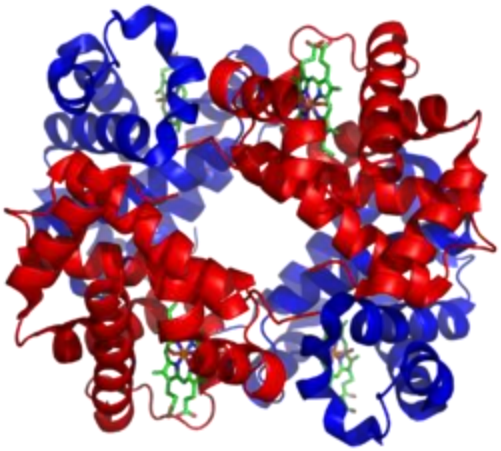
- **Why did I include human myoglobin?**

Following the procedures we used for the cross-species myoglobin tree, build a phylogenetic tree for the human globins. (I recommend using the same settings you used for the earlier tree to facilitate comparisons between them.) Here is the tree I produced:



- **What can you infer about the expansion of the (hemo)globin family from this tree?**
- **If you were to build a (hemo)globin tree from another species, would it have the same structure? Would it even have the same members?**

97% of adult hemoglobin complexes consist of two subunits of hemoglobin alpha (red) and two subunits of hemoglobin beta (blue) associated as a heterotetramer ($\alpha_2\beta_2$):



(image from wikipedia.org)

The remaining 3% of adult hemoglobin complexes contain other combinations of subunits; the biological significance of these atypical complexes is not well understood.

In a developing human fetus, the dominant hemoglobin complex contains two alpha subunits and two gamma subunits ($\alpha_2\gamma_2$). The $\alpha_2\gamma_2$ configuration binds oxygen more tightly than adult hemoglobin, allowing the fetus to "steal" O_2 molecules from mom's circulating blood.

- **Based on the tree above, which other hemoglobin subunit seems most likely to bind to hemoglobin alpha? Explain your reasoning.**
- **What other form of functional genomic data (beyond sequence similarity) could you collect in support of your answer above?**

Part 3: Inferring the origin of the hemoglobin family

Select one of the human hemoglobin proteins from Part 2 and copy/paste its sequence into the myoglobin sequence dataset from Part 1. Repeat the tree-building procedure from Part 1.

- **Where does the hemoglobin fall in the resulting tree?**
- **What do you infer from this about the timing of hemoglobin evolution?**

Extension: Bootstrapping

Bootstrapping is a statistical technique for assessing confidence in an analysis. This method involves resampling your data with replacement to make new datasets, repeating your analysis on the new datasets, and comparing the new ("bootstrapped") results to the original results. The word "bootstrapping" comes from the phrase "pull yourself up by your bootstraps" (relevant here since we're assessing confidence using a single dataset and not independent datasets.)

Bootstrapping is often used to estimate a confidence interval around a sample measurement. For example, imagine we have two lists of numbers of length *N* with a Spearman correlation coefficient of 0.40. We create 1,000 new pairs of lists of length *N* by randomly sampling (x, y) pairs from the original lists. We compute a correlation coefficient for each of the newly-sampled pairs. If 95% of these "bootstrapped" correlations fall between 0.15 and 0.65, then we would be fairly confident that the true value of the correlation exceeds 0.

While the sampling used in bootstrapping is reminiscent of permutation testing, there is a critical difference: bootstrapping maintains associations between paired data, while permutation testing breaks those associations (to see how often "such a strong result" occurs at random).

Bootstrapping is used in phylogenetic analysis to measure our confidence in the branches of a tree. The original MSA is resampled by picking random columns (with replacement) and concatenating those columns to make a new MSA of equal length. The tree is then rebuilt from the resampled multiple alignment. This procedure is repeated many times. Let's say that for four sequences A, B, C, and D the original tree had the following topology: ((A,B),C,D). However, if only 50% of the resampled trees have this topology, we would say that our bootstrap confidence in the pairing (A,B) is low. In other words, A may well be closer to C or D—the data are unclear. Several of the options on the Phylogeny Workflow website use bootstrap values as measures of support for branches in the resulting tree (red numbers). Other methods (such as maximum likelihood) estimate these values directly when building the tree.

Examine one of the multiple alignment files from today's activity:

- **Using Python, how might you load and store the aligned sequences from an MSA?**
- **How could you create a new MSA by resampling the sequences stored above?**