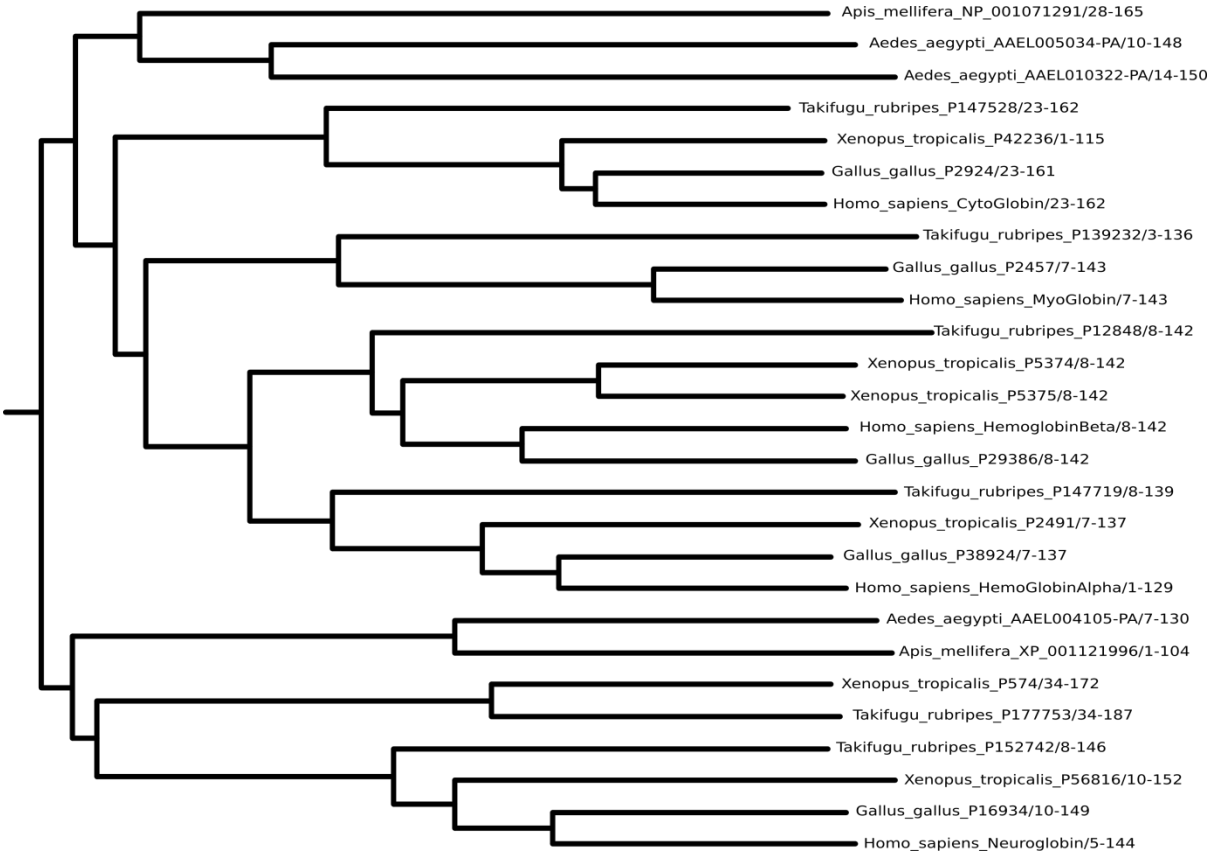
BioSB questions day 2:

Question 1: Orthology and paralogy in the globin family



a) Inspect the above tree of globin proteins. Annotate all internal nodes in terms of duplications and speciations using pen or pencil.
b) Mark / annotate in the tree where you think it is likely that two gene loss events occurred.

Question 2: Orthology and protein names

The human proteins APOO and APOOL and the yeast proteins YGR235C and YNL100W are all part of the MICOS complex that is involved in making cristae in mitochondria. These proteins are all homologs of each other.

a) In the unified nomenclature, APOOL and YNL100W have been named MIC27, while YGR235C has been named MIC26. A recent paper discovered that APOO is also part of MICOS and named it MIC26. Do you agree with this?

Hint: Naming proteins implies orthology. Is there evidence for this? Or are the

phylogenetic relationships different? Make an alignment + phylogeny. Collect the proteins, put them in an alignment program (see yesterdays instructions, or just use clustal) and create a phylogeny.

b) Some gene duplications are actually the result of complete genome duplications. Two well known genome duplications are the one at the root of the vertebrates and the one in an ancestor of *Saccharomyces cerevisiae*. Can you find evidence that the human proteins resulted from a genome duplication.

Hint: Make a tree, e.g. in Blast, and examine which species have two homologs and which ones only one. For the human proteins you can also look up the tree in Treefam.

Question 3, pushing the outside of the envelope of homology detection

Mic12 (Aim5) is a short, rapidly evolving protein and simple, pairwise sequence homology detection have not been able to find orthologs of this protein outside of the fungi.

a) Can you find a human homolog?

Hint: You will have to use all the homology detection tools that you can lay your hands on…like HHpred. Make sure you select enough iterations in HHpred (at least 5).

b) What, if anything, is conserved in this protein family?

Hint: in HHpred the amino acids that are strongly conserved in a protein family are written in capitals. When they are conserved between families, the capitals are connected with a vertical line. Alternatively you can put the sequences in JACKHMMER, iterate the search and examine the sequence logo.

c) Is there corroborating evidence that the human candidate protein is indeed located in mitochondria?

Hint: Check "protein atlas" for localization data of human proteins. Another name for LOC125988 is C19orf70 Also entering a gene name in pubmed can be useful.

d) Can you find a potential homolog in Arabidopsis thaliana, and if so, can you find corroborating evidence for the location of this protein in the inner mitochondrial membrane?

Hint: check the gene name of the Arabidopsis homolog (the AT... name) in entrez

(http://www.ncbi.nlm.nih.gov/) In order to find a homolog in Arabidopsis you can use a combined alignment of the metazoan and fungal homologs (hhalign) and run that with hhpred against the Arabidopsis genome. You can also download this alignment from the BioSB course site: QIL1-AIM5.aln

Question 4, Complications of a gene duplication and accompanied loss of sequence similarity.

a) The human gene ATP6AP1 is known to function as an assembly factor of the vacuolar ATPase (V-ATPase). Does it have yeast homolog(s). If there are multiple homologs, do they align with different parts of the human gene? If so, are they the result of a fission event? Or did something else happen?

Hint: predicting the transmembrane topology of the homologs gives you some intuition here. But you can also just examine the length of the proteins.

b) If you find borderline cases of homology (e.g. the E-value is between 1 and 0.001), can you think of a strategy to confirm them?

Hint: Realize that homology is transitive: if A is homologous to B and B is homologous to C then A is homologous to C. Maybe there is another fungal species in HHpred that has a homolog to ATP6AP1 that you can use as "intermediate" sequence (notice that there in principle are no intermediate sequences in sequence space. Nevertheless, in this case where the gene duplication has been followed by accelerated sequence evolution in parts of the proteins, it might help).

c)The human protein ATP6AP1 is actually cut by a protease at Arg-225, resulting in two proteins that are attached. How does that cleavage site compare with your results?