

Tuesday, 10th September 2002

1. Find cytochrome c protein sequences for at least five of the following: human, horse, pig dog, rabbit, kangaroo, chicken, penguin, rattlesnake, tuna, dogfish, fruit fly (*Drosophila*), baker's yeast (*Saccharomyces cerevisiae*), wheat. Edit these to be a consistent length (i.e., around 104 residues).
2. Perform global multiple sequencing alignments for these, and for the beta globin sequence set prepared yesterday:

- use both **Pileup** and **ClustalW**;
- limit input to 5 or 6 sequences;
- view both the text and graphical outputs.

3. Build trees of these, too:

- use **Proudest**, and examine the distance matrix generated;
- use UPGMA and neighbor joining options;
- pipe output to **Growtree** and view the output.

4. Make a profile (**Profilemake**) for each set (using the **Pileup** output). Save the consensus from each profile.
5. Examine the output! What are 'B', 'Z'? What does the bottom row of the profile represent? How do these matrices differ? How do they encode the information in the original alignment?
6. Repeat your search of the database from yesterday using:
 - a) each profile (**ProfileSearch**)
 - b) each consensus
 - c) the most distantly related sequence from each set.

7. Exploration of HMMs: The web sites below are servers for algorithms that either find genes in DNA sequence or create profile HMM protein domain models. Use one or more to locate the exons in one or both of two sequences fragments. The fragments are in two files, "t.seq" and "g.seq" in the [Group Home] area of your account directory. **NB: You must first copy the files to your own subdirectory in order to use them!** It may be better to work in groups for this. As these sites are often slow to return results, you will probably need to return to this problem later in the week. Once you are confident that coding region has been accurately identified, you may wish it perform database searches with the translated and/or

perform appropriate annotation using some of the tools available at sites outside of ANGIS.

Chris Burge (Genscan)	ccr-081.mit.edu/chris
Sean R. Eddy (HMMER)	www.genetics.wustl.edu/eddy
Anders Krogh (SAM)	genome.cbs.dtu.dk/krogh
David Haussler	www.cse.ucsc.edu/~haussler/index.html
Genie Web Server	www.cse.ucsc.edu/~dkulp/cgi-bin/genie
ORNL Grail Form	grail.lsd.ornl.gov/Grail-1.3/
GeneScan	ccr-081.mit.edu/Genscan.html

8. A short fragment of DNA from a mouse “shotgun sequencing” project has been placed in the [Group Home] area of your account directory as “unknown00.seq.”
NB: You must first copy this file to your own subdirectory in order to use it!
- a) Find any DNA sequences that have significant similarity to some or all of unknown00.seq
 - b) Find the protein sequence most similar to the translated sequences.
 - c) Build a profile using any apparent homologues and research the protein database. What new, similar sequences do you find using this strategy?
 - d) Draw a diagram of what you believe to be the most likely domain structure for a protein encoded by this gene. (You will need to use tools outside of ANGIS for this.)