

Laboratory Worksheet, Tuesday, Feb. 19.

I. The Actin-HSP link. From two weeks ago, you should have the core families from the Pfam database for their actin and HS70 families saved. (If you do not, please do that today.) As an exercise in multiple sequence alignment, I would like you to try to “shuffle these families together”, that is, I would like you to take several members from each family and first, make pairwise sequence alignments, first within the family and then between the two families. Then take a small sample from each of the two families and make one family of, say, six or eight sequences, and align them using a multiple sequence aligner. Which three or four sequences would you take for your sample? Please write up your results for the week after the break (Tuesday, March 4). I will collect these.

II. Codon Usage Bias Project. Recall that we agreed to focus upon the three species (from among those considered by Karlin-Mrázek): *E. coli*, *M. tuberculosis* and *P. horikoshii*. By Tuesday, March 4, you should at least have a program which calculates the codon usage bias functions described by Karlin-Mrázek for a single file and saves this data to a file. We should agree to do this for the HS proteins (do a couple of those for the background) and for at least one non-HS protein, from each of the three species. You can choose the genes to be analyzed. I will not collect these programs yet.

The goal will be to automate this comparison process, but for now, we haven't touched upon enough Perl to be able to download files automatically from a database. the final goal of this project would be to be able to do this and, in principle, be able to make such computations on a fairly large scale.

III. An old project, not forgotten. Recall that we looked at the sequence from Entrez with accession number **CAC01034** from *Leishmania major*. Please run a search of this segment against the protein portion of GenBank and against Pfam. Is there any information newer than that first posted by the sequence's discoverers?

IV. Perl Resources. While gearing up for our Perl exercises, we will need to streamline programs by using subroutines. I will try to get the second and third hours of Perl tutorial scheduled today, but in the meantime, it would be useful for you to look at the archive of Perl modules available at the **Comprehensive Perl Archive Network** (<http://www.cpan.org/>). Look especially at the BioPerl package of modules. The class resource page now has a link to the search page for CPAN. You should take a couple of minutes to explore this feature.

V. Some extra tools. There are some interesting alignment “editing” tools available on the web. The two I have in mind at the moment are: motif logos which produce visual summaries of the entropic profile of a piece of multiple sequence alignment. One is available at <http://www-lmmb.ncifcrf.gov/toms/sequencelogo.html>. Actually, examples and documentation are available there, and the logo maker is available via a link. CINEMA helps

display information in a multiple sequence alignment visually. It can be accessed through <http://www.biochem.ucl.ac.uk/bsm/dbbrowser/CINEMA2.02/kit.html>. This requires installation, however, and I will try to get it installed by the week after the break.