

Homework 4

L519: Bioinformatics: Theory and Applications
School of Informatics
Indiana University – Bloomington
March 3, 2003

Due date: March 10, 2003

Working Directory : Use /gpfs/L519/hw4 on IBM SP. /gpfs on IBM SP and /bigscr/bioinfo/ on Solar are temporary files systems and files on them may be lost in case of the system failure. *Be sure to make copies of your work* somewhere else.

Instruction : This will use both hw2 and hw3. There are about 3,500 COG families on the directory /bioinfo/data/COG/. Using the COG family used for your hw3 (you can choose a different one if you want), you will build a hidden Markov model (HMM) and search for matches in the Agrobacterium genome (which is not included in COG) using the model.

- Select a family with more than 40 seqs, say COGxxxx under the directory /bioinfo/data/COG. You can use the file /bioinfo/data/COG/COGnoseqs.
- Make symbolic links to the files, COGxxxx, AE007869.faa, AE007870.faa, AE007871.faa, AE007872.faa, See the example below (*do not copy them*).

```
$ cd my-working-directory
$ ln -s /bioinfo/data/genbank/genomes/Bacteria/Agrobacterium_fumefaciens/AE007869.faa .
```

- Perform a multiple sequence alignment and build a HMM as below.

```
$ clustalw COGxxxx > any-file-name
$ hmmbuild filename-for-model COGxxxx.aln
$ hmmsearch filename-for-model AE007869.faa > filename-search-output.69
$ hmmsearch filename-for-model AE007870.faa > filename-search-output.70
$ hmmsearch filename-for-model AE007871.faa > filename-search-output.71
$ hmmsearch filename-for-model AE007872.faa > filename-search-output.72
```

- Perform again the above experiment with a refined model. In the above, you build a HMM using the whole multiple sequence alignment. However, as in the case of the phylogenetic tree construction, it will result in better performance when only conserved regions in the multiple alignment are used.

Data : You can use data on /bioinfo/data/COG and /bioinfo/data/genbank/genomes/Bacteria/Agrobacterium on IBM SP.

Hint : Try AE007871.faa, that have only 198 proteins, on command line for your debugging, and then submit your job using the l1submit command for search against all four .faa files.

Make sure to use different machines while you debug your programs. I suggest students in a different group project use a different machine (say aries01 for the microarray group, aries02 for the protein-protein interaction group, etc).

Extract conserved regions in a multiple seq alignment is not trivial. Use whatever method you know for that. If you do not know such a method, use the procedure below. Instead of extracting conserved regions from a multiple sequence alignment, just define a conserved region (start and end positions) for each sequence, and use your hw2 (borrow from others if you are a non-major student) to extract the regions and perform multiple sequence alignment.

Discussion Forum : hw4 on oncourse.iu.edu

If the forum is not set, please email me immediately or call me.

Directory and submission :

- Create a directory `/gpfs/L519/hw4/yournetwork-id`
- Put the three scripts on `/gpfs/L519/hw4/yournetwork-id`
- Put the result from the program execution on `/gpfs/L519/hw4/yournetwork-id` with a name `your-network-id.hw4.result`
- Create the README file where a line under RUN will execute your program without any modification.
- Submit a printed version of your document and put also an electronic version of the document on `/gpfs/L519/hw4/yournetwork-id` with a name `your-network-id.hw4.report`. *The document should contain*
 - the name of COG family and the annotation (a single line) of the family from the COG database,
 - the result from searching with the first HMM: a list of match sequence ids and their annotations if any,
 - the result from searching with the refined HMM: a list of match sequence ids and their annotations if any, and
 - a brief discussion on the difference between two searches if any.