

Homework 3

L519: Bioinformatics: Theory and Applications
School of Informatics
Indiana University – Bloomington
Feb. 20, 2003

Due date: March 3, 2003

Working Directory : We are using `/bigscr/bioinfo/L519/hw3` on Solar. `/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.

Intruccion : There are about 3,500 COG families on the directory `/bioinfo/COG/`.

- 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`.
- For each sequence s_i in `COGxxxx`,
 - Perform a Blast search (`blastp`) with a query s_i against `COGxxxx`.
 - By parsing the Blast result, count how many matches are found with $\text{Evalue} < e^{-1}$. Let's denote the set of matches $M_i = \{s_{i_1}, \dots, s_{i_j}\}$.
- Write a document that summarizes your searches as below. Include only sequence-id.
 - *A table of records* each of which is a triplet $(s_i, M_i, \text{percent})$, i.e., (query-sequence-id, a set of matches, percent of matched seqs).
 - In most cases, “the percent of matched seqs” is much below 1, which means that we miss seqs in the family with a Blast search. Select two sequences, say s_{max} and s_{min} , where the percent for s_{max} is the maximum and the percent for s_{min} is the minimum. Write down the sequence ids for s_{max} and s_{min} .
 - For the two sequences, s_{max} and s_{min} , generate multiple sequence alignments for each seq, say run `ClustalW` with two set of sequences (each set containing the query and matches). Look for any significantly different regions by comparing the two multiple sequence alignments. The comparison can be done either manually or automatically (if you know any good methods to compare two multiple sequence alignments). *Include in your document* the two multiple seq alignments and the regions where the two multiple seq alignments differs, if any.
 - *Write a brief paragraph* discussing why a Blast search miss some seqs that belongs to the same family.

Data : You can use data on `/bioinfo/data/COG` on IBM SP and Solar.

Discussion Forum : hw3 on `oncourse.iu.edu`

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

Directory and submission :

- Create a directory `/bigscr/bioinfo/L519/hw3/yournetwork-id`
- Put the three scripts on `/bigscr/bioinfo/L519/hw3/yournetwork-id`
- Put the result from the program execution on `/bigscr/bioinfo/L519/hw3/yournetwork-id` with a name `your-network-id.hw3.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 `/bigscr/bioinfo/L519/hw3/yournetwork-id` with a name `your-network-id.hw3.report`.