

Algorithms in Bioinformatics II, SS2003

Assignment sheet # 3

Daniel Huson

May 12, 2003

1 χ^2 test (3 points)

Consider an indicator variable I that can take on the values 0 or 1 and a second variable X that can take on the values A, C, G, T . Assume that we are given a contingency table obtained from a training set that tells us the number $f_i(H)$ of data sets for which $I = i$ and $X = H$, for $i \in \{0, 1\}$ and $H \in \{A, C, G, T\}$:

| I | X | | | |
|-----|----------|----------|----------|----------|
| 0 | $f_0(A)$ | $f_0(C)$ | $f_0(G)$ | $f_0(T)$ |
| 1 | $f_1(A)$ | $f_1(C)$ | $f_1(G)$ | $f_1(T)$ |

Describe in detail how to use χ^2 statistics to determine whether the variable X depends on I or not.

The remainder of your assignment this week is to build and use a profile HMM to search for hemoglobin sequences.

2 Obtain the initial sequences (1 point)

Here are the accession numbers of an initial set of hemoglobin sequences: P01922, P01958, P02023, P02062, P02185, P02240, P04252 and P22431. Please grab the protein sequences from GenBank and put them into a fasta file called `sequences.fasta`.

3 Build a multiple alignment (1 point)

Find a web-server for ClustalW. Using ClustalW, generate a multiple sequence alignment of the data set in a file called `sequences.aln`.

4 Build a profile HMM (2 points)

Find a web-server running the HMM package HMMER. Using HMMER, build a profile HMM for the given set of sequences and report it in a file `sequences.hmm`. (If you have difficulties running a web server on the output file generated by ClustalW due to format problems, then please reformat the file `sequences.aln` into fasta format.)

5 Description of the constructed HMM (1 point)

Give a detailed description of the profile HMM generated by HMMER.

6 Searching for sequences using a profile HMM (2 points)

Use the constructed HMM to search the *SwissProt release* protein database for a number of additional hemoglobin sequences and summarize your results.

Hint You may need to run HMMER on different web-servers. I found one web server that could be used to generate an HMM whose description was easily downloadable. Unfortunately, I was not able to get the HMM to run properly from this web-site. I found a second website that was happy to build and run an HMM for me and returned the results to me within a few minutes. However, this second website didn't give me access to the actual HMM that was built.

Assignments due: **Monday, May 19, 10am**