Bioinformatics Homework

http://biochem158.stanford.edu/bioinformatics.html

Homework Assignment

1. Select a protein of interest to you from UniProt/SwissProt database whose function is well known and well characterized. Obtain the FASTA format of the protein and the Gene Ontology terms associated with your protein.

2) Search your protein for homology using BLAST to search UniProt/SwissProt. Do NOT search all of UniProt. Please report two or three hits which are both statistically and biologically significant. Also report two or three hits which you think are neither statistically nor biologically significant. If your protein family is very large, you may have to ask BLAST to return more hits to find statistically insignificant hits. The NCBI BLAST will let you view up to 20,000 hits if need be.

3) Search your protein for motifs with the MyHits Motif Scan Query. Be sure to INCLUDE Prosite patterns, frequent patterns, Prosite Profiles, more profiles and Pfam Local HMMs in your search. Please send me the MyHits that you think are biologically significant and at least 1 or 2 hits which you think are not biologically significant. You judge biological significance by comparing the function of the pattern hits with the gene ontology terms associated with your query. Be sure to include high frequency patterns in order to be able to discover some biologically insignificant hits. The high frequency patterns are NOT protein functions, but instead represent possible protein modification sites. The profiles and HMM hits will have expectation values associated with them which will help you determine significance.

3) Search your protein for blocks using the InterPro database. Please send me a few of the InterPro domains hits you think are significant and any which you think are not biologically significant. InterPro scan shows you the predicted gene ontology terms for your query. Are these correct?

Statistical vs. Biological Significance

Biological Significance

First, for each search (Prosite, InterPro and BLAST), I would like you to report some biologically significance hits and describe why you think they are significant biologically. Also report some biologically insignificant hits. One judges biological significance by determining if the query and the hit share any gene ontology terms. Proteins may be similar because they are the same protein in different organisms, because they are both members of the same protein functional family, because they are both members of the same structural family or because they just share one domain or motif (like an ATP binding motif). Tell me at what level of biological significance of each hit you report.

Statistical significance and expectation values for the BLAST search.

Statistical significance is determined by the expectation value that gives you a measure of how likely this finding is based on pure chance. A finding with an E-value of 1 or greater is not significant because it could occur by pure chance. A finding with an E-value less than 10^{-3} (one chance in a thousand) is generally considered statistically significant (unless of course you are doing a 1,000 searches!). So the lower the expectation value, the more significant the finding. Findings between 10^{-3} and 1 are in the so-called twilight zone and require some further analysis or experiments to determine their validity.

Statistical vs. Biological Significance (cont.)

InterPro

Unlike most of the other methods, InterPro sets a very high level of significance for a finding before it will report it. This means that you will often not find any statistically insignificant hits for this particular search.

Biological Significance

In order to determine biological significance you must read the biological properties of your protein and the biological properties of your findings. The findings may be significant because the finding defines a very closely related protein family (opsins for example) or a very broad family (G-coupled protein receptors or 7-transmembrane proteins) or a common structure (protein fold) or a specific function (retinal binding site) or a very specific catalytic activity. You should describe in words the level of the biological significance.

Statistical vs. Biological Significance (cont.)

Myhits

MyHits will return patterns as well as profile hits from its Prosite Database. You will notice that patterns do not have scores or E-values associated with them so there is no easy way to judge statistical significance. With pattern findings you are left only with judging biological significance. The best way to do this is to read the documentation on each hit you find. Also none of the frequent patterns from Prosite are statistically significant. The frequent patterns do NOT represent functions in your query. Instead they represent potential protein modification sites in your query. You can judge their biological significance by looking for Protein Modification sites in the UniProt entry for your query. You can also do a Google Scholar search for your "query protein name" AND "name of modification site"

BLAST

If you do not have any insignificant hits from the BLAST search, it means that your protein family is very large and you have to ask BLAST to return more results using the Advanced Options at the bottom of the form. Only when you see hits with E-values > 0.001 do you have insignificant findings.

Copying Website Output to Homework Doc

Copying sequence alignments to your homework email message or document

When copying sequence alignments to either an email message or a document, the font often gets changed to a variable spaced font (one where each letter has a different width). In order to keep the sequence alignments aligned, you must select the sequence alignment lines (and their sequence numbering lines as well) and change them back to a monospaced font like Monaco or Courier, fonts in which each letter has exactly the same width.

Copying graphics information to your message or document.

Graphics information on your findings from the web sites can be copies to the clipboard and then pasted into your message or document using special graphics capture keystrokes. For the Macintosh, Command-Shift-3 will copy a selected region of the screen to the clipboard and Command-shift-4 will copy the entire screen to the clipboard. On the PC, Function (Fn key)+ (PRT Sc) print screen key will copy the screen to the clipboard.