Homework Assignment 3

http://biochem218.stanford.edu/03Homework.pdf

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. Briefly discuss what the protein does and why you chose it.
2. Search your protein for similar sequences using the BLAST method on the UniProt site. Please search ONLY the UniProt/SwissProt section of the database. Please report two or three hits which are both statistically and biologically significant. One judges biological significance by looking to see if protein shares any gene ontology terms with your query sequence. One judges statistical significance by looking for expectation values less than 0.001. 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.
3. Search your protein for motifs with the Prosite Motif Scan Query. Please send the Prosite hits you think are biologically significant and at least 1 or 2 hits which you think are not biologically significant. Be sure to include high frequency patterns in order to be able to discover some biologically insignificant hits.
4. Search your protein for motifs using the InterPro database. Please report a few of the InterPro domains hits you think are significant and any hits which you think are not biologically significant. InterPro scan shows you the predicted gene ontology terms for your query. Are these correct?
5. Send your results and conclusions to bioc218-spr1314-staff@lists.stanford.edu. Please include sufficient output from your analyses (copy and paste or screen dumps) to support each of your answers/conclusions.

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 which 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 biologically insignificant hits for this particular search.

Biological Significance

In order to determine biological significance you must read the gene ontology descriptors of your protein and the gene ontology terms of your hits or matches. 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)

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. You may have to use the NCBI BLAST site that permits returning 20,000 results. Make sure you are only searching the SwissProt section of the database. Only when you see hits with E-values > 0.001 do you have insignificant findings.

Statistical vs. Biological Significance (cont)

Prosite

The Prosite patterns do not have E-values associated with them so there is no easy way to judge statistical significance. With Prosite you can only judge biological significance. 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 Amino Acid Modification sites table 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”. Another method would be to do a MeSH search in PubMed for “query protein name” AND “Mesh term for your modification”[MeSH]. Either of these methods would allow you to validate any frequent Prosite Protein modification hit.

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 copied to the clipboard and then pasted into your message or document using special graphics capture key strokes. 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.