

Hints for Bioinformatics Homework

<http://biochem118.stanford.edu/bioinformatics.html>

Homework Assignment

- 1) Select a protein from [OMIM](#) concerning the disease of interest to you.
- 2) Search your protein for motifs with the MyHits Motif Scan Query. Be sure to Include Prosite Patterns, Prosite Frequent Patterns, Prosite Profiles, Prefiles, Pfam HMMSs (local Models) in your search. Please send me the MyHits you think are biologically significant and at least 1 or 2 hits which you think are not statistically or biologically significant. Please note that only the Profiles have expectation values. The Patterns do not have a measure of statistical 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 at least 1 or 2 hits which you think are not statistically or biologically significant. Please note that the default graphic output of InterPro does not list expectation values. You must switch to the Tabular view to obtain the statistical significance.
- 4) Search your protein for homology using the BLAST method. 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.



Statistical vs. Biological Significance

- Assignment
- First, for each search (MyHits, Blocks, InterPro and pBLAST), I would like you to report some significance hits and describe why you think they are significant both statistically and biologically; also report some statistically insignificant hits (and why) and are any of your statistically insignificant hits, still significant biologically). To remind you what I said in class: a statistically significant find in the database search is always biologically significant, but a biologically significant result in the search is not necessarily always statistically significant.
- Statistical significance and expectation values.
- 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 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
 - If you ask MyHits to return PATTERNs as well as motifs, you will notice that PATTERNs do not have 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. Also none of the Frequent patterns from MyHits are statistically significant.
- 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 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.