Bioinformatics Resources
Exercise Sheet 3
discussed on May 19th and May 22nd

1. a) What is the longest human protein in UniProtKB? Describe the query you used and give the name and length of the protein.
   b) Assuming that this is a sufficient identifier, query UniProtKB for the gene name of the protein you found. In how many other organisms can you find a hit? You might want to switch to the UNIX command line for parts of this task.

2. a) From a study you are given the following GI numbers 515374, 538202, 5881724, 114591. What are the respective (primary) UniProtKB accession numbers? For this part, use an approach, that would also work for a 1000 GI numbers. You can assume there is always a perfect 1-to-1 mapping.
   b) What organism are these proteins from? What is their subcellular localisation? What is the identifier of the KEGG pathway that all of these proteins are involved in?
   c) For the protein that maps to GI number 114591: What is the evidence code for the two sequence conflicts? How can you tell this from the XML? Is this a manual or automatic annotation?

3. The following UniProt accession numbers describe an important protein from different organisms: P01308, P30410, P67974, P04667, P67970. Using a UniProtKB service, verify whether all annotated disulfide bonds are conserved. Which part of the sequence, that is specifically annotated by UniProtKB, is missing in one of the entries?

Good luck,
J. Reeb & L. Richter