Bioinformatics Resources
Exercise Sheet 2
due on May 15th, 9 a.m.
Total number of points: 14

Task 1
How many sequences are in the current release of UniProtKB/TrEMBL? In relation to UniProtKB/SwissProt? Plot the sizes and the relations over the last ten years, one data point per year. (2P)

Task 2
a) What is the longest human protein sequence in UniProtKB? Describe the query you used and give the name and length of the protein. (2P)
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? Hint: You might want to switch to the UNIX command line for parts of this task. (4P)

Task 3
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. (2P)
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? (2P)

Task 4
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? (2P)

Good luck,
J. Reeb & L. Richter