Assignment 1 / Biometry and bioinformatics I / 2013

The file gene.txt includes a DNA-sequence from a gene.

Answer the following questions and submit your report to Moodle (link in course webpage, will be ready to receive submissions 16. September ->) Monday 23. September at the latest.

Your answers need not be long, a few sentences are enough, just to show that you can find information from relevant sources.

1. What is the source of this sequence (the accession number(s)) ?

2. By using the OMIM-database, give a short (just a few sentences) description about this gene.

3. What is the most recent scientific publication concerning this gene? (Use PubMed-database)

4. Pick up the corresponding sequence from 15 other animals (including human.) In addition to the nucleotide database which we have used during the tutorial sessions, you can also use the RefSeq-database.

5. Explain briefly what is the main difference between the nucleotide database and RefSeq database.

6. Align the sequences. You can use Clustal (installed in class C128), or http://mafft.cbrc.jp/alignment/server/ (or some other facility), or you can use the alignment facility (and also the data-mining facility) in MEGA5 (http://www.megasoftware.net/) by first learning how to do this by taking a “Walk through MEGA” (http://www.megasoftware.net/tutorial.php) .

7. Inspect the alignment by using your own eyes and brains (cf. the tutorial example: what kind of gaps are reasonable) and write about possible errors, or write that there are no errors. You don´t have to start editing the alignment manually; it is enough that you realize that there might be something to be edited. (In case you want to start leaning, by yourself, also this kind of matters now, the program Bioedit has been installed to C128 machines.)

   Include the aligned FASTA-file in your assignment report. This means that you include a text-file (look at the tutorial example: there is a) the original FASTA-file, and next to it there is b) the aligned one and now you should include in your report the file which is like b)).

8. By using MEGA5, calculate the p-distance matrix from your sequences. To show that you understand what this matrix means, pick up a couple of examples from the matrix: what is the most similar species pair and their p-distance, what is the most dissimilar species pair and their p-distance.

9. By using MEGA5, construct the UPGMA-tree (use the p-distance option and do not pay attention to any kind of model selection etc. options). Copy-paste the tree in your report and write a couple of sentences as a description of it (i.e. what kind clusters you can notice.)