BIOMETRY AND BIOINFORMATICS I, exam 21.10 – 3.11.14

Submit your answers 3.11.14 at the latest, to course Moodle-area.

All answers must be included in one document and their order must be: 1.-4. All pictures, or other "not-text components" must be included within the answers – NOT as separate documents. As some questions are numbered like (for example) 1.1, 1.2, etc., your answers should also be numbered like this.

You can choose the document format for your answer document, but you are not allowed to use a zipping. In addition to the answer document, you should submit a text file as indicated in question 1.

You must work alone and any indications of collaboration may lead to: http://blogs.helsinki.fi/alakopsaa/processing-a-suspicion-of-cheating/?lang=en

_____

Questions 1 and 2 worth 8 points, questions 4 and 5 are 4 points. => max 24 points.

12 points: grade 1
14 points: grade 2
16 points: grade 3
18 points: grade 4
20 points: grade 5

You will receive personal response about your answers when the exam period is over, i.e.during two first weeks of November.

The extra assignments (5 cr -> ->10 cr) will appear in course webpage after 3.11. There is no deadline for doing extra assignments, and when you submit, say, one extra assignment, your credits will be extended from 5 cr to 6 cr, and if you submit another extra assignment, for example, one month later, then your credits will be extended from 6 cr to 7 cr. Etc.

_____

Question 1.

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

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

1.1 What is the source of this sequence (the accession number(s))

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

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

1.4 Pick up the corresponding sequence from 10 other animals (including human.)

1.5 Align the sequences. There are several possibilities: You can use Clustal (installed in class C128), or http://mafft.cbrc.jp/alignment/server/ , 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) . You can also use some other aligning facility.

1.6 Inspect the alignment by using your own eyes and brains (cf. the course assignment: what kind of gaps are reasonable) and write about possible aligning 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.

Include the aligned FASTA-file in your exam answers as a separate document. This means that you include a text-file: look at the course

assignment 1: there are two text documents including the data, (a) Initial dataset for assignment 1.txt and (b) Initial dataset for assignment 1 aligned.txt. You should give your data like in (b).

1.7 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: the most similar species pair and their p-distance, what is the most dissimilar species pair and their p-distance. Tell, what they are, and what are their p-distances.

1.8 By using MEGA5, construct the neighbor-joining tree by using the p-distances and another tree by using the Jukes-Cantor model. Include the trees in your answer and write a couple of sentences as a description: what kind of clusters you can notice, what does the confidence analysis (bootstrapping) tell, and do the p-distance and Jukes-Cantor model usage result in differences in the trees.

Question 2.

Datafiles “bact_genedata1.txt” and “bact_genedata2.txt” include data from two protein coding genes from bacteria. There are five species: *Streptococcus pneumoniae, Streptococcus mitis, Streptococcus oralis, Streptococcus agalactiae and Streptococcus thermophilus.*
All those sequences, which are named by numbers or numbers+letters belong to the species *Streptococcus pneumoniae.* They are it´s different types, so called serotypes. Sequences named by multiple numbers or multiple numer+letter combinations (for example 25F_25A_38): sequences from those types are identical, and combined for convenience. *Streptococcus agalactiae* and *Streptococcus thermophilus* are also represented by different types which are named as you can notice by looking at the data.
Construct neighbor-joining trees from both genes. Use the Jukes-Cantor model and bootstrapping for confidence evaluation.
Show both trees in your answer sheet.

Producing correct trees gives you 2 points out of 8 points from this question.

The rest, 6 points, comes from interpretations, i.e. you should write about the results:

- Describe the structures of phylogenetic trees which you have produced, i.e. how the sequences (which come from different species and, on the other hand, are different types from a given species) are related to each other, what kind of clusters they form.
- You should pay attention to two levels: between-species level and within-species level.
- Describe also how do the two genes differ: if you find that they produce different kind of phylogenetic trees, explain what are the differences – and again: regarding between-species and within-species levels.

You can use either “original tree” or “consensus tree” , it does not matter very much, but use the same option for your both trees.
You can notice that the aligned datafiles have not been completely edited after alignment, but don´t care about that - does not affect the results, and do not assume that your – maybe differing results from the two genes - could result from this.
You may also notice that not all completely identical sequences have been combined. This is also a minor issue and has no effect on results which you should explain.

Question 3.

A student is frustrated because he should complete an assignment right now, and he tells you that he wants to get a better phylogenetic tree than he now has: some of it´s bootstrapping values are still very small, and he has been running already 10 000 replicates. Help him. Give some good advise.

Question 4.

Consider one site from a stretch of nucleotide sequence, from which there is data from six OTUs. The data at this site is:
OTU1 has A, OTU2 has A, OTU3 has T, OTU4 has C, OTU5 has C and OTU G.
How many different rooted topologies there can be, in total, considering this single site?
Consider one possible rooted topology and infer it´s HTU´s by using parsimony criteria. (HTU = hypothetical taxonomic unit, an inner node, including the root node.). You can either draw the tree, or give it as NEWICK-format and explain what are the HTUs.