

# Exam in DD2396 Bioinformatics

## June 2, 2010

Lecturer: Lars Arvestad

No aids beyond writing equipment are accepted. **Write clearly! Justify your answers!** Leave room for comments during grading. Please use only **one side of each paper**.

A passing grade (E) is awarded at 15 points, 20 points are required for grade C, and 25 points for grade A. See the course web page for an explanation as to how this relates to course grades! Each successful homework will give one bonus point for part one, but the maximum points for part one is 15.

Exam results will be emailed to the participants.

*Note:* Part 2 will only be graded if part 1 has been awarded at least 10 points (including bonus).

**Good luck!**

### Part 1

*Please note:* on this published exam, the numbering has been fixed. The class-room version had lost the number on question 2, and all subsequent question numbers were off by one.

1. Answer TRUE or FALSE for the following claims. You get a half point for each correct answer. (2p)
  - (a) Multialignment programs cannot guarantee that they find the most optimal alignment.
  - (b) “Sequence masking” is a gene prediction method.
  - (c) Secondary structure prediction is easier for transmembrane proteins than for proteins in general.
  - (d) We can determine homology using synteny data.
2.
  - (a) What is the guiding principle for choosing phylogenetic trees under *parsimony*? (1p)
  - (b) Why is it problematic to build phylogenetic trees from short sequences? (1p)
  - (a) In a *semiglobal*, or *ends-free*, alignment, flanking columns containing indels are disregarded, i.e., they get score 0.  
Why are semiglobal alignments often to be preferred when aligning data from shotgun sequencing or EST sequences? (1p)
  - (b) What do we call the type of alignment that Blast is using? (1p)
3.
  - (a) How do you define *genome coverage*? (1p)
  - (b) What is a *contig*? (1p)
  - (c) What is meant by “paired-end reads”? (1p)
4. Draw an approximate sequence logo for the motif described by the alignment in Figure 1. (2p)

```
GATGGA
TATAAA
GATAGA
TATAAA
GATAGT
TATAAT
GATAGT
TATAAT
```

Figure 1: A motif for question 4.

5. (a) Figure 2 shows the beginning of a typical Blast report. What is the most significant hit? How many significant hits do we have? Justify your answer! (2p)
- (b) When you use Blast to compare coding DNA with coding DNA, it is usually recommended to instruct Blast to translate (in all reading frames) the DNA and compare amino acid sequences instead of nucleotide sequences. Why? (2p)

```

Query= p0825.6.C1 nseq=5
      (826 letters)

Database: sprot
        144,731 sequences; 53,363,726 total letters

Searching.....done

Sequences producing significant alignments:

Score      E
(bits)     Value

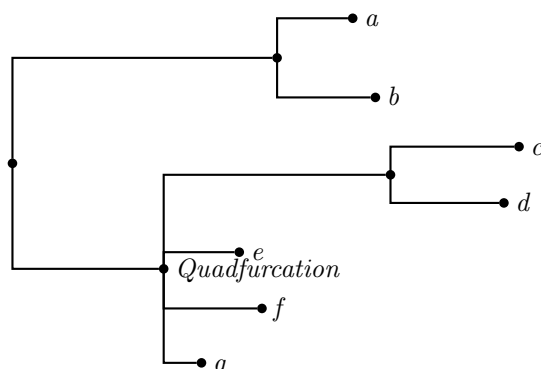
APOD_HUMAN (P05090) Apolipoprotein D precursor (Apo-D) (ApoD)      82  1e-15
APOD_CAVPO (P51909) Apolipoprotein D precursor (Apo-D) (ApoD)      81  2e-15
APOD_MOUSE (P51910) Apolipoprotein D precursor (Apo-D) (ApoD)      79  1e-14
APOD_RABIT (P37153) Apolipoprotein D precursor (Apo-D) (ApoD)      78  2e-14
APOD_RAT (P23593) Apolipoprotein D precursor (Apo-D) (ApoD)      73  7e-13
ICYB_MANSE (Q00630) Insecticyanin B form precursor (Blue bil...    55  1e-07
ICYA_MANSE (P00305) Insecticyanin A form (Blue biliprotein) ...    49  1e-05
BBP_PIEBR (P09464) Bilin-binding protein precursor (BBP)            46  1e-04
ERBP_RAT (P06911) Epididymal-retinoic acid binding protein p...    41  0.004
RET2_ONCMY (P24775) Plasma retinol-binding protein II (PRBP-II)    39  0.010
RETB_XENLA (P06172) Plasma retinol-binding protein precursor...    39  0.018
RETB_CHICK (P41263) Plasma retinol-binding protein precursor...    39  0.018
LAZA_SCHAM (P49291) Lazarillo protein precursor                    38  0.023
RET1_ONCMY (P24774) Plasma retinol-binding protein I (PRBP-I)     38  0.030
BLC_VIBCH (Q08790) Outer membrane lipoprotein blc precursor ...   37  0.052
LACB_FELCA (P33687) Beta-lactoglobulin I                           37  0.052
PURP_CHICK (P08938) Purpurin precursor                             37  0.052
CRA2_HOMGA (P80007) Crustacyanin A2 subunit                        35  0.26
VE2_HPVO8 (P06422) Regulatory protein E2                           35  0.26
CRC1_HOMGA (P80029) Crustacyanin C1 subunit                        34  0.34
CRA1_HOMGA (P58989) Crustacyanin A1 subunit                        34  0.34
AMBP_PLEPL (P36992) AMBP protein precursor [Contains: Alpha...    34  0.44
LACC_FELCA (P33688) Beta-lactoglobulin III                         34  0.44
RETB_HORSE (Q28369) Plasma retinol-binding protein precursor...    33  0.57
RETB_PIG (P27485) Plasma retinol-binding protein precursor (...    33  0.75
RETB_HUMAN (P02753) Plasma retinol-binding protein precursor...    32  1.7
PGHD_RAT (P22057) Prostaglandin-H2 D-isomerase precursor (EC...   32  2.2
VE2_HPVS5B (P26545) Regulatory protein E2                          31  2.8
VE2_HPVO5 (P06921) Regulatory protein E2                           31  2.8
RETB_BOVIN (P18902) Plasma retinol-binding protein (PRBP) (RBP)   31  2.8
VE2_HPVS6 (P50809) Regulatory protein E2                           31  2.8

```

Figur 2: Resultatet av en Blast-sökning. Används i fråga 5.

## Part 2

6. A bootstrap analysis often results in trees with multifurcations (i.e., some nodes have more than two children in a rooted tree, see the “quadfurcation” below). What do they mean? (2p)



7. Please outline the general structure of a simple HMM-based *ab initio* gene-finder for eukaryots. Please describe the features that your HMM contains. (3p)
8. Suppose we generate 1000 *completely random* protein sequences of length 1000 and build a Blast database from them. Then, we generate a completely random protein sequence to be used as a Blast query against the random database. What would you expect (roughly, in orders of magnitude) the best  $E$  value to be? Explain! (2p)
9. Suggest a method for defining protein families and discuss at least one advantage and one disadvantage with the method. You can devise your own method or describe a well-known method. The input to your method is a large set of protein sequences. (3p)
10. There are programs available, for example “GBlocks” and “TrimAl”, that reduce multialignments by removing columns that are considered to interfere with phylogenetic estimation. The idea is that some columns probably introduce noise and if we can identify them, we would get more reliable results.
  - (a) What do you think these softwares look for when searching for columns to remove? What could the characteristics be for chosen columns? (2p)
  - (b) It is popular to use GBlocks and similar programs in automatic analysis pipelines, i.e., when several analysis steps are chained and performed on many datasets. When researchers are only working on a single dataset, these methods are rarely (if ever) used. Why is that? (1p)
  - (c) Obviously, some column removals will be wrong. Can you suggest a case when a removal could worsen a result? No points for obvious cases such as “if we remove it, we get the wrong tree”. Give a specific example, with or without sample data. (1p)
  - (d) Suppose you are asked to make a large-scale evaluation of GBlocks and TrimAl in order to determine which of the two programs are better and if they are at all worthwhile using. How would you do it? (1p)