Exam I&B Integration: Sequence Analysis
(Lecturer A.P. Gultyaev)
Monday December 19, 2016, 10:00-13:00

Exam consists of 10 questions, maximum scores are indicated. The grade is the sum of points obtained for all answers. The total grade is equal to: 0.8 × (exam grade) + 0.2 × (course assignments' grade). All answers require some (short) argumentation. Both English and Dutch may be used. Good luck / succes !

1. (0.5 points). The following exon annotation is given for a mRNA containing 8 exons:

```
exon    1..266
        267..447
        448..624
        625..699
        700..810
        811..864
        865..900
        901..2489
```

The protein-coding sequence is annotated as follows:

```
CDS     239-955
```

A functional motif has been identified at the amino acid positions 161-170 of the encoded proten. Which of the mRNA exons encode(s) this motif ?

2. (0.5 points). The algorithms for sequence database similarity search like BLAST and FASTA exploit the strategy of identifying "word" similarities in query and subject sequences. Initial steps of these algorithms use the thresholds of word lengths, adjusted for the optimal algorithm performance. Where are these thresholds higher: in the algorithms designed for nucleic acids or in those for proteins ?

3. (1.0 points). Two amino acid sequences have been globally aligned by a dynamic programming algorithm. The alignment includes a number of insertions and deletions in both sequences. How to determine whether this alignment is statistically significant ?

4. (1.0 points). A BLAST result for some nucleotide sequence query using BLAST database "nr/nt" (BLAST database containing entries from various species) yielded a human (*Homo sapiens*) sequence as a hit with relatively low E-value ("Expect"): say, E = 3e-13. Of course, a BLAST search with the same query in the same database, but with a constraint by organism name (*Homo sapiens*), yields the same hit. What is the most likely E-value in this case:
(a) E = 3e-13;
(b) E = 3e-12;
(c) E = 2e-12;
(d) E = 2e-14.
Explain your answer.

5. (0.5 points). A sequence database similarity search using a DNA sequence as a query in standard BLASTN program (searching a nucleotide database) has yielded no results. However, using the same query in BLASTX program (searching protein database using a translated nucleotide query) yielded a number of significant hits. How can this be explained and what can be concluded from this result ?

6. **(1.0 points)**. In addition to protein sequence database similarity hits, the program BLASTP (protein-protein sequence similarity database search) also reports the presence of known sequence motifs in the protein query. Frequently the results of this motif search are yielded considerably faster than sequence hits are produced. Why ?

7. **(1.0 points)**. The progressive multiple alignment algorithms are so-called "greedy" algorithms. Why do they get this label ? Give an example of strategy that could diminish the greedy nature of progressive multiple alignment.

8. **(2.0 points)**. Below a position-specific score matrix (PSSM) for a transcription factor binding site is given:

```
A [ 27   0   1 27 27 20 ]
C [  0   0   9  0  0  0 ]
G [  0   0   0  0  0  1 ]
T [  0 27 17  0  0  6 ]
```

Try to suggest how to change the DNA sequence given below, using only two substitutions, so that to obtain two binding sites, located on two complementary DNA strands. The binding sites should have scores higher than 130.

```
1 - GGATGAATCG CGCGTTGATG - 20
```

9. **(0.5 points)**. Searching for long reading frames can be efficiently used for gene finding in prokaryotic genomes, but almost useless in eukaryotic genomes. Why ?

10. **(2.0 points)**. Below a fragment of dynamic programming matrix calculated by an algorithm for global pairwise sequence alignment is shown for the 3'ends of two nucleotide sequences: ...TTTCG-3' (seqX) and ...TTGCTCCG-3' (seqY). The scoring rules are as follows: match +1; mismatch -1; gap penalty -2 for each nucleotide in a gap.

```
... T   T   T   C   G   3' (seqX)
T  14  12  10   8   6
T  12   .   .   .   .
G  10   .   .   .   .
C   8   .   .   .   .
T   6   .   .   .   .
C   4   .   .   .   .
C   2   .   .   .   .
G   0   .   .   .   .
3'
(seqY)
```

(a) Fill in the missing numbers in the matrix (dots).
(b) What is the score of the global alignment ?
(c) Show the sequence alignment.