Biological Sequence Analysis (Spring 2015)

Exercise 1

Thu 15.1, 10-12, B222

Choose any 5 assignments from below (each assignment gives 1 point, 5 points is maximum for each week).

1. **Sequence databases — paralogs**

   Find human hemoglobin paralogs alpha, beta, gamma and delta in NCBI sequence database: `http://www.ncbi.nlm.nih.gov/sites/entrez?db=Nucleotide`. Compare sequences using some aligner.

2. **Sequence databases — orthologs**

   Find insulin orthologs from human and mouse in NCBI sequence database. Compare sequences using some aligner.

3. **Alignment scores**

   Give a scoring scheme that yields score 6 for the alignment below.

   ```
   CAGCA-CGTACAACAGCTACCA
   CATCACCG--C--CA--TAG-A
   ```

4. **Justification for score matrices.**

   Prove that Kullback-Leibler divergence (see lecture slides) is always non-negative.

5. **PWMs and PSSMs**

   Some binding sites for hematopoietic transcription factor GATA-1 from *H. sapiens* are listed below:

   ```
   AGATAA
   TGATAA
   AGATAG
   TGATAG
   TGATCA
   TTATCA
   ```

   Compute the consensus sequence, positional weight matrix (PWM), and position-specific scoring matrix (PSSM) for the sites as described at the lecture (using pseudocounts for the latter). Compute also the sequence logo heights for the letters at each position.

6. **Searching with palindrome PSSM.**

   Modify the example given at lecture `http://www.cs.helsinki.fi/u/vmakinen/bsa15/pssm.py` to work with palindrome PSSMs like `AGAACAnnnTGTTCT`.

7. **Motif discovery and statistical significance.**

   Given a set of N promoter sequences each of length L, an *exact motif finding* problem can be formulated as the task of finding k-mers that occur in n out of N promoter sequences (at least once in each) and have small probability of occurring that many times in a random set of sequences following the same distribution as the promoter sequences.

   Let C_w denote the number of promoter sequences containing k-mer $w = w_1 w_2 \cdots w_k$.

   a) Derive an estimate for the expected value of C_w assuming the background follows the i.i.d. model.

   b) Why C_w divided by its expected value does not give a good ranking for reporting the statistically most significant k-mer motifs?

   c) Find out what kind of different rankings (statistical tests) are used in this kind of contexts. What do you need to know about the distribution of values C_w to use them?

8. **Generating DNA sequences with higher-order Markov chains.**

   Write a program (e.g. in python) to read the k-th order distribution of a given DNA sequence (for given k), and to generate a new sequence of the same length simulating the same distribution.