Introduction to bioinformatics, Autumn 2007, Exercise 4

3.10.2007

1. (a) (Chapter 7, Exercise 4) For the sequences I = `GCATCGGC` and J = `CCATCGCCATCG`, find matching 4-words shared by I and J. Do this by making tables $L_W(I)$ and $L_W(J)$.

   (b) Describe an efficient algorithm for finding occurences of the k-word B in the string A given the table $L_W(A)$.

2. (Chapter 7, Exercise 7) For I = `TTGGAATACCATC` and J = `GGCATAATGCACCCC`, make dot matrices for the k-tuple hits for $k = 1, 2, 3$.

3. (Chapter 7, Exercise 9) For I = `GTATCGGCGC` and J = `CGGTTCGTATCGTCG`, make a 2-word list for J. Compute diagonal common word sums for I and J using the algorithm presented in lectures and in course book (Computational Example 7.1).

4. (Chapter 7, Exercise 10) To search J = `CGGTTCGTATCGTCG` for matches to `TTCG` within one mismatch, first make a list of all possible matches. How many matches are there within a single mismatch neighborhood of TTCG? [Hint: There is one exact pattern, and there are $3 \times 4 = 12$ single-mismatch patterns.

5. Download a DNA sequence from `http://www.cs.helsinki.fi/mbi/courses/07-08/itb/data/seq1.faa` and use it as the query sequence in the following tasks.

   (a) Run FASTA nucleotide at EBI (`http://www.ebi.ac.uk`) against EMBL Coding Sequence database. Choose "interactive" as the parameter Results. Otherwise use default parameters.

   (b) Run nucleotide BLAST at NCBI (`http://www.ncbi.nlm.nih.gov`) against NCBI RefSeq database (choose Reference mRNA sequences database). Otherwise use default parameters.

   Explain the contents of the result page from both programs in your own words. How many matches did you get? How similar were the best matches to the query sequence? Did you receive same results from both programs?