Genomics and Bioinformatics

Examination - Week 7

October 30, 2012

Question 1 - Sequence Alignment

Linear gap penalty

Using the following scoring: a match is worth 2 points, a mismatch is penalized -1, and a gap is penalized -2 points, find all optimal alignments of the sequences
`ATTCCGTTA`, `ATCGA`
by the Needleman-Wunsch algorithm.

Affine gap penalty

Calculate the scores of the previous alignments using a gap opening penalty of -2 points and a gap extension penalty of -1. Are the previous alignments still equivalent?

Question 2 - Phylogenetic trees

The *INS* gene (insulin peptide precursor) is shared among many vertebrates.

The table below provides BLAST scores of pairwise alignments for six *INS* family proteins from four different species,

1. **hs_ins** from *Homo sapiens*, i.e. human,
2. **xt_ins** from *Xenopus tropicalis*, i.e. frog,
3. **rn_ins1** and **rn_ins2** from *Rattus norvegicus*, i.e. rat,
4. **mm_ins1** and **mm_ins2** from *Mus musculus*, i.e. mouse.

	hs_ins	xt_ins	rn_ins1	rn_ins2	mm_ins1	mm_ins2
hs_ins	-	414	499	612	589	421
xt_ins	444	-	111	125	221	91
rn_ins1	546	178	-	667	976	690
rn_ins2	411	129	800	-	702	899
mm_ins1	466	241	1112	680	-	630
mm_ins2	666	113	800	932	780	-

1. Based on the table above, draw the gene tree for the six *INS* proteins (**mm_ins1**, **mm_ins2**, **rn_ins1** and **rn_ins2**, **xt_ins**, **hs_ins**). Indicate speciation and duplication events on the graph.

2. Using this tree, give examples of:

   - an orthologous pair,
   - a paralogous pair in the same species, and
   - a paralogous pair in different species.

Question 3 - Genome Assembly

Consider the following reads: TGCATGC, CCATGCA, CCATGTC, ATGCCAT

1. Construct the overlap graph based on the above reads, using an overlap size of 4 bases. Draw a path that goes through every *vertex* (Hamiltonian path), and write the corresponding contig.

2. Make a list S_4 of all unique 4-mers (9 elements) and the list S_5 of all (non-unique) 5-mers (12 elements).

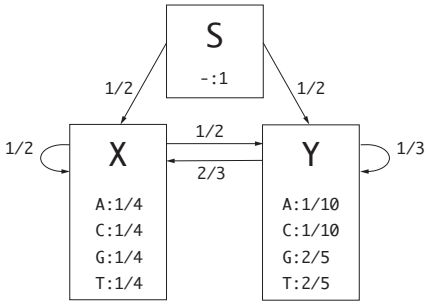
3. Construct the De Bruijn graph with S_4 as vertex set and S_5 as edge set.

4. Add one edge to make this graph Eulerian and find two Eulerian paths. Write down the corresponding contigs and indicate which of them is compatible with the full reads.

Question 4 - Hidden Markov Model

Consider the Hidden Markov Model represented in the figure below. The hidden states are S, X and Y, with emitted symbols $-$, A, C, G and T. The non-zero emission probabilities are specified under the corresponding hidden states.

Use the Viterbi algorithm to fill the scoring table and then use backtracking to find the most probable hidden state sequence associated with the observed sequence "CGA":



	-	C	G	A
S				
X				
Y				