Genomics and Bioinformatics

Examination - Week 7

November 1, 2011

Question 1

Consider the following reads: AATGCAT, GCATGCA, TGCAATG, AATGCGA

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 an eulerian path. Write down the corresponding contig.

Question 2

In this question you will compute the optimal global alignment between these two sequences using the Needleman-Wunsch algorithm.

Sequence 1: `ACGTATAGGC`

Sequence 2: `ACTAAGC`

1. Sequence alignment
   First, find the alignment that has the highest score using the classical version of the Needleman-Wunsch algorithm. In this version a match is worth +1 point. A mismatch is penalized -1 point. An insertion/deletion (gap) is penalized -2 point.
2. Affine gap penalty
   Secondly, calculate the score of the above alignment with a gap opening penalty of -2 and an gap extension penalty of -1. Using these new rules, is there a different alignment with a higher score ?

Hint

The gap extension penalty is counted for every gap, the opening penalty only for the first gap in a consecutive stretch. For example three consecutive gaps cost $-5 = -2 + 3 * (-1)$.

For this second part, you don't need to fill a dynamic algorithm table, just compute the score and try to improve it.

Question 3

In this question you will build a profile HMM based on the following multiple sequence alignment:

```
ACT
AAT
ACC
TCT
TCC
```

1. Convert each column of this multiple alignment to nucleotide frequencies. Set three hidden 'match' states M_1, M_2, and M_3 which will emit with the frequencies of the corresponding columns in the alignment.

2. Set four hidden 'insert' states I_0, I_1, I_2, and I_3 which all emit with background frequencies (1/4 each base).

3. Transitions are as follows: I_n has transitions to itself (probability 20%) and to M_{n+1} (except I_3), M_n has transitions to M_{n+1} (probability 80%) and I_n, except M_3 which only has a transition to I_3.

4. Draw the HMM diagram, write the transition matrix $\mathcal{M}$ and the emission matrix $\mathcal{E}$. Fill in the missing transition probabilities so that the matrix is consistent.

5. Complete the table below using Viterbi's algorithm and give the most likely sequence of hidden states for the test sequence `GACAT`

   Remark: All probabilities have been multiplied by 5 to simplify the calculations, the results are unchanged.

	-	G	A	C	A	T
I_0	1	5/4	25/16	125/64	625/256	3125/1024
M_1	0	0	15	0	375/16	625/32
I_1	0	0	0	75/4	375/16	1875/32
M_2	0	0	0	240	75	0
I_2	0	0	0	0	300	
M_3	0	0	0	0		
I_3	0	0	0	0		

6. Convert the resulting sequence of hidden states into an alignment containing all the sequences of the multiple alignment above and this last test sequence.

Question 4

Circadian rhythms (often referred to as the "body clock") are ubiquitous in most life forms. The "period" family of genes is an essential component of this clock.

The table below provides BLAST scores of pairwise alignments for six period family proteins from three different species,

1. **dm_per** from *Drosophila melanogaster*, i.e. fruit fly,
2. **dr_per2** and **per3** from *Danio rerio*, i.e. zebra fish,
3. **mm_per1**, **mm_per2** and **mm_per3** from *Mus musculus*, i.e. mouse.

	dm_per	dr_per2	dr_per3	mm_per1	mm_per2	mm_per3
dm_per	-	116	125	129	123	121
dr_per2	114	-	661	887	947	497
dr_per3	123	644	-	667	666	699
mm_per1	126	745	572	-	640	80
mm_per2	126	984	680	832	-	455
mm_per3	124	499	709	537	488	-

1. Based on the table above, draw the gene tree for the six period proteins (dm_per, dr_per2, dr_per3, mm_per1, mm_per2, mm_per3).
2. Using this tree, give examples of:
   - an orthologous pair,
   - a paralogous pair in the same species, and
   - a paralogous pair in different species.