Genomics and Bioinformatics

Examination - Week 7

October 29, 2013

Question 1 - Genome Assembly

Consider the following reads: ATGATGC, TGCATGA, ATGCCAT, CCATGCA

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 (10 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. Is this graph Eulerian?
5. Find two Eulerian paths, write down the corresponding contigs, and indicate which of them is incompatible with the full reads.

Question 2 - Sequence Alignment

Linear gap penalty

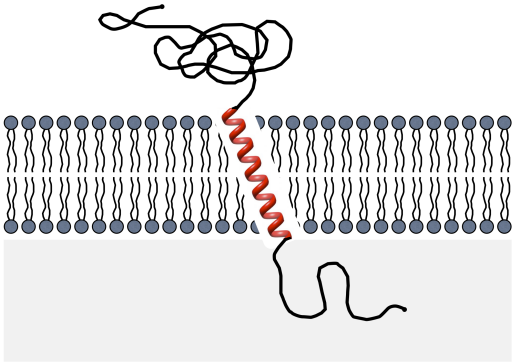
Using the following scoring: a match is worth 2 points, a mismatch is penalized -1, and a gap is penalized -2 points, find 3 optimal alignments of the sequences
`AACTTTG`, `ACCTG`
by the Needleman-Wunsch algorithm. What is their score ?

Affine gap penalty

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

Question 3 - Hidden Markov Model

A transmembrane protein is a protein with one or more domains going from one side of the plasmic membrane through to the other side. As a consequence, it contains extracellular, intracellular and transmembrane domains. Very often the transmembrane domains are structures known as “alpha-helices”. We want to find them using a Hidden Markov Model.



The propensity of different amino-acids to be part of an alpha-helix is well-known and scored in the table below. For this exercise we assume that

- Transmembrane domains are alpha-helices of 20 amino-acids on average.
- Amino-acids are equally distributed in the other domains.
- Intracellular domains are 200 amino-acids long on average and extracellular domains, 400 amino-acids.

Table 1: Frequency of amino-acids within helical domains:

A	R	N	D	C	E	Q	G	H	I
8%	7%	4%	4%	4%	6%	6%	3%	5%	6%

L	K	M	F	P	S	T	W	Y	V
7%	6%	7%	5%	**0.3%**	5%	4%	5%	5%	5%

1. Propose a simple Hidden Markov Model to predict transmembrane domains from a given amino-acids sequence (draw a diagram).
2. What are the emission probabilities?
3. Based on the length of the respective domains, give an estimate of the transition probabilities. Write the transition matrix.
4. Here is the sequence of a dummy transmembrane protein. How would you calculate the probability of observing this sequence, given the model? (Do not calculate it).

   NGAKTTL

5. On the same protein, we underlined the transmembrane domain. What is the probability of observing this sequence?

   NG<u>AK</u>TTL

6. *Bonus:* how would you adapt the model (only the diagram) if moreover you know that an alpha-helix is defined as hydrogen bonds linking groups of 4 amino-acids?

Question 4 - Homology

Toll-like receptors (TLRs) are a class of proteins involved in the innate immune system. They all share similarity to the *Drosophila* protein Toll and are found in vertebrate and non-vertebrate species. Some of their motifs can be found in plants and bacteria, suggesting that they are components of an ancient immune system.

The table below provides BLAST scores of pairwise alignments for six TLR family proteins (from 3 different species) and the *Drosophila* Toll protein,

1. **TLR4_BT** from *Bos taurus*, i.e. bovine,
2. **TOLL_DM** from *Drosophila melanogaster*, i.e. fruit fly,
3. **TLR3_HS** and **TLR4_HS** from *Homo sapiens*, i.e. human,
4. **TLR3_MM** and **TLR4_MM** from *Mus musculus*, i.e. mouse.

	TLR4_BT	TOLL_DM	TLR3_HS	TLR4_HS	TLR3_MM	TLR4_MM
TLR4_BT	-	460	177	1263	205	1061
TOLL_DM	412	-	188	358	150	404
TLR3_HS	169	255	-	160	1406	117
TLR4_HS	1238	375	166	-	206	1087
TLR3_MM	169	120	1422	235	-	186
TLR4_MM	1041	413	133	1074	102	-

1. Based on the table above, draw the gene tree for the six proteins. Indicate speciation and duplication events on the graph.
2. Using this tree, give examples of:
   - one orthologous pair,
   - one paralogous pair in the same species, and
   - one paralogous pair in different species.