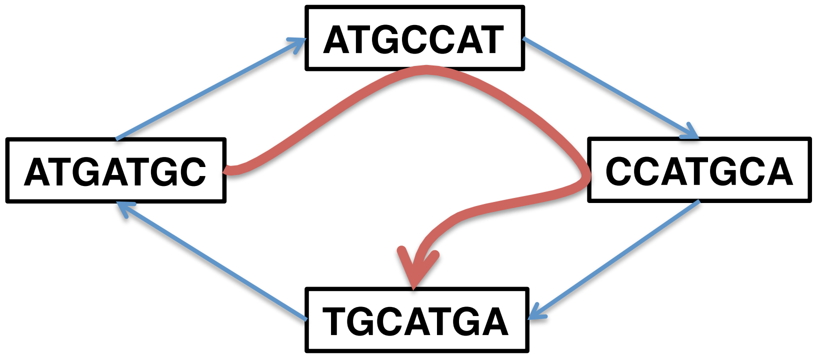
Genomics and Bioinformatics

Exam correction

October 29, 2013

Question 1 - Genome Assembly



1. A Hamiltonian path follows the red arrow and generates the following contig:
ATGATGCCATGCATGA

2. The 4-mers and 5-mers sets are:

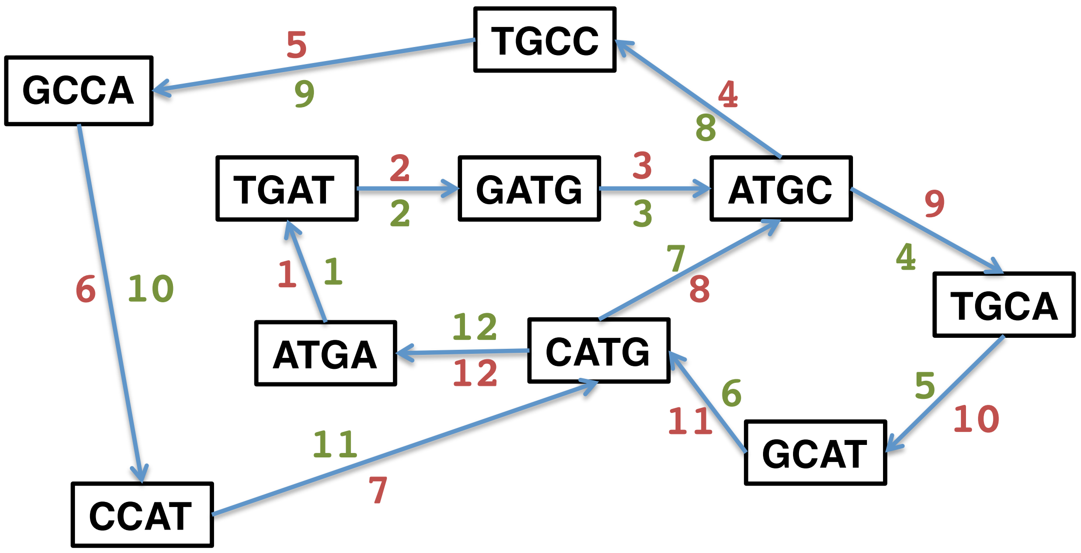
$$\begin{aligned} S_4 &= \{\mathbf{ATGA}, \mathbf{TGAT}, \mathbf{GATG}, \mathbf{ATGC}, \mathbf{TGCC}, \\ &\quad \mathbf{GCCA}, \mathbf{CCAT}, \mathbf{CATG}, \mathbf{TGCA}, \mathbf{GCAT}\}. \\ S_5 &= \{\mathbf{ATGAT}, \mathbf{TGATG}, \mathbf{GATGC}, \mathbf{ATGCC}, \mathbf{TGCCA}, \mathbf{GCCAT}, \\ &\quad \mathbf{CCATG}, \mathbf{CATGC}, \mathbf{ATGCA}, \mathbf{TGCAT}, \mathbf{GCATG}, \mathbf{CATGA}\}. \end{aligned}$$

3. The de Bruijn graph is as below:

4. This graph is Eulerian: for every vertex the number of incoming edges is always equal to the number of outgoing edges.

5. One eulerian path follows the edges in the order of the red numbers and generates the contig
ATGATGCCATGCATGA
and the green numbering leads to
ATGATGCATGCCATGA

The second contig could not have generated the reads **CCATGCA** and **TGCATGA**.



Question 2 - Sequence Alignment

Linear gap penalty

	-	A	A	C	T	T	T	G
-	0	-2	-4	-6	-8	-10	-12	-14
A	-2	2	0	-2	-4	-6	-8	-10
C	-4	0	1	2	0	-2	-4	-6
C	-6	-2	-1	3	1	-1	-3	-5
T	-8	-4	-3	1	5	3	1	-1
G	-10	-6	-5	-1	0	4	2	3

The nine optimal alignments have score = 3 and are given as follows:

```
AACTTTG  AACTTTG  AACTTTG  AACTTTG  AACTTTG  AACTTTG  AACTTTG  AACTTTG  AACTTTG
ACCT--G  ACC--TG  ACC-T-G  -ACCT-G  -AC-CTG  -ACC-TG  A-CCT-G  A-C-CTG  A-CC-TG
```

Affine gap penalty

The first two alignments have 1 gap opening, so their score is

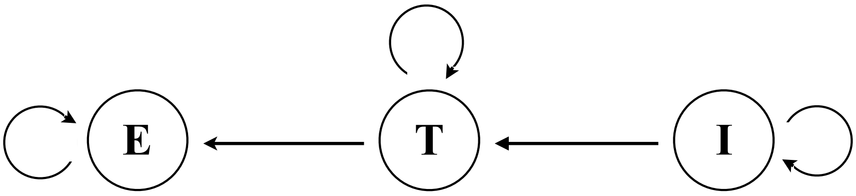
$$3 - 1 \times 2 \; = \; 1 \; .$$

All the others have 2 gap openings, so a score of

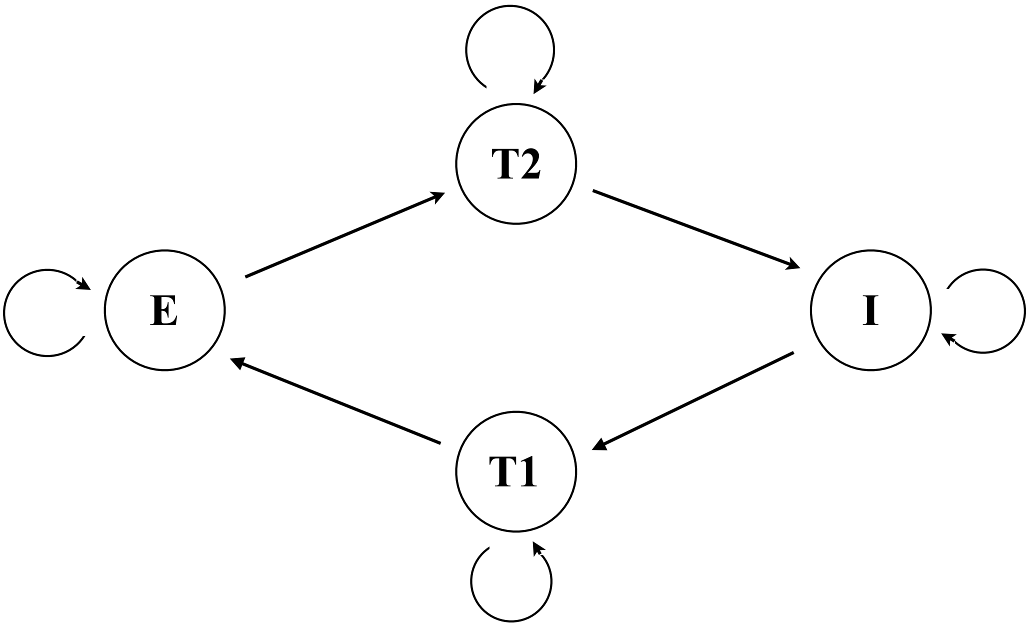
$$3 - 2 \times 2 \; = - \, 1 \; .$$

Question 3 - Hidden Markov Model

1. Let us call T a transmembrane domain, E an extracellular domain, I an intracelular domain: these are the hidden states. The simplest diagram is as follows:



but this doesn't allow more than one transmembrane domain. One could add reverse arrows, but then it would allow $I \to T \to I$ transitions. To accomodate several transmembrane domains we can use two identical states $T1$ and $T2$:



There may be other models, but we give the solution only for these two.

2. The emissions from the T state are given by the helical propensity table. In other domains, since all amino-acids are equiprobable, each has a frequency $1/20 = 5\%$.

3. As we saw in an earlier exercise, the transition probabilities can be set as the inverse of the expected segment length. In this case, there are about 20 amino-acids in a transmembrane domain, so we can set the probability of going out of T as 1/20.

   If we use two T states, $T1 \to E$ and $T2 \to I$ transitions both have probability 1/20. Outgoing probabilities must sum to 1, so the transition $T \to T$ is 19/20.

   In the E state, the outgoing probability is 1/400 (therefore $E \to E$ is 399/400) and similarly for I we have 1/200 and 199/200. This yields the following transition matrices:

$$\begin{array}{c} E \\ T_2 \\ I \\ T_1 \end{array} \begin{pmatrix} 399/400 & 1/400 & 0 & 0 \\ 0 & 19/20 & 1/20 & 0 \\ 0 & 0 & 199/200 & 1/200 \\ 1/20 & 0 & 0 & 19/20 \end{pmatrix} \quad \text{or} \quad \begin{array}{c} E \\ T \\ I \end{array} \begin{pmatrix} 1 & 0 & 0 \\ 1/20 & 19/20 & 0 \\ 0 & 1/200 & 199/200 \end{pmatrix} .$$

   One can also use 1/201, 200/201, etc. as seen in the exercise.

4. Use the Forward algorithm to compute the total probability of the observed sequence.

5. If O_n is the n-th observation and S_n is the n-th hidden state, we can write

$$P(O|S) = P(O_1O_2O_3 \ldots |S_1S_2S_3 \ldots) = P(O_1|S_1) \cdot P(O_2|S_2) \cdot P(O_3|S_3) \cdot \ldots .$$

   Emissions are equally probable in states E and I: $P(O_1|S_1) = P(N|E) = 5\%$ *etc.*, and in the T state we have $P(O_3|S_3) = P(A|T) = 8\%$ and $P(O_4|S_4) = P(K|T) = 6\%$. Finally:

$$P(O|S) = (5 \cdot 5 \cdot 8 \cdot 6 \cdot 5 \cdot 5 \cdot 5)/100^7 = \mathbf{1.5 \cdot 10^{-9}} .$$

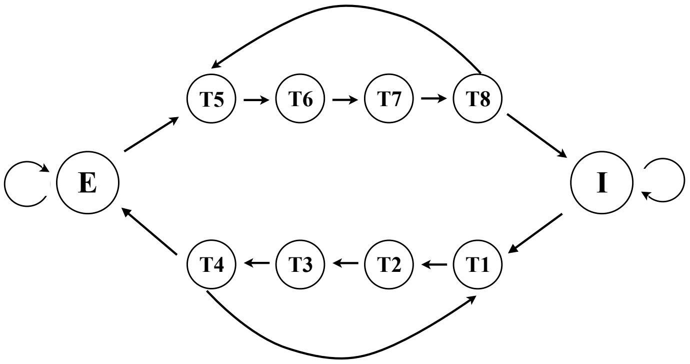
   Alternatively: One could also understand the question as finding $P(O, S)$ instead of $P(O|S)$, in which case we have $P(O, S) = P(O|S) \cdot P(S)$. We need

$$P(S) = P(S_1|S) \cdot P(S_2|S) ... = P(\text{initial state is } S_1) \cdot P(S_2|S_1) \cdot P(S_3|S_2) \cdot ...$$

   which are the transition probabilities. Multiply that with the $P(O|S)$ from above. This version requires to choose in which domain you start (E or I). For instance, starting from I and using the unidirectional 3-states model, the known hidden sequence is $IITTEEE$ and we have:

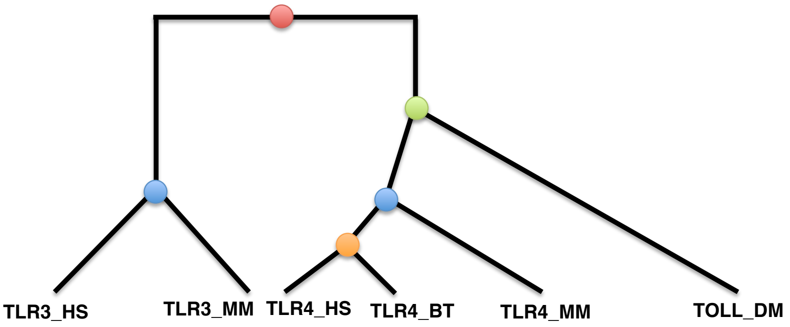
$$P(O, S) = 1.5 \cdot 10^{-9} \cdot (1 \cdot \frac{199}{200} \cdot \frac{1}{200} \cdot \frac{19}{20} \cdot \frac{1}{20} \cdot \frac{399}{400} \cdot \frac{399}{400})$$

6. If transmembrane domains always contain a multiple of 4 amino-acids, one can extend the model as in the figure below. Another solution would be to emit quadruplets of amino-acids when in state T, each with its specific probability.



Question 4 - Homology

1. The tree below is compatible with the similarity scores. The red dot is a gene duplication, the green, blue and orange dots denote three speciation events.



2. Orthologous pairs:

3. **TLR3_HS** and **TLR3_MM**, any combination of **TLR4_MM**, **TLR4_HS**, **TLR4_BT** and **TOLL_DM**.

4. Paralogous pairs in the same species:
   TLR3_HS and **TLR4_HS**, **TLR3_MM** and **TLR4_MM**.

5. Paralogous pairs in different species:
   TLR3_HS or **TLR3_MM** with any other protein in a different species.