

Suggested solutions: Lab 5

Task 1 – Pair-wise sequence alignment

In the case of ties, this solution chooses a (mis)match over an indel. This means that it is only possible to read out *one* optimal path from the arrows (even if there exist more than one optimal path).

		S _{i,j}									
			G	G	A	C	G	T	A	C	G
S _{i,j}		0	1	2	3	4	5	6	7	8	9
0	0	0	→ -1	→ -2	→ -3	→ -4	→ -5	→ -6	→ -7	→ -8	→ -9
T	1	↓ -1	↘ -1	↘ -2	↘ -3	↘ -4	↘ -5	↘ -4	→ -5	→ -6	→ -7
A	2	↓ -2	↘ -2	↘ -2	↘ -1	→ -2	→ -3	→ -4	↘ -3	→ -4	→ -5
C	3	↓ -3	↘ -3	↘ -3	↓ -2	↘ 0	→ -1	→ -2	→ -3	↘ -2	→ -3
G	4	↓ -4	↘ -2	↘ -2	→ -3	↓ -1	↘ 1	→ 0	→ -1	→ -2	↘ -1
G	5	↓ -5	↘ -3	↘ -1	→ -2	↓ -2	↘ 0	↘ 0	↘ -1	↘ -2	↘ -1
G	6	↓ -6	↘ -4	↘ -2	↘ -2	↘ -3	↘ -1	↘ -1	↘ -1	↘ -2	↘ -1
T	7	↓ -7	↓ -5	↓ -3	↓ -3	↘ -3	↓ -2	↘ 0	→ -1	↘ -2	↓ -2
A	8	↓ -8	↓ -6	↓ -4	↘ -2	→ -3	↓ -3	↓ -1	↘ 1	→ 0	→ -1
T	9	↓ -9	↓ -7	↓ -5	↓ -3	↘ -3	↘ -4	↘ -2	↓ 0	↘ 0	→ -1

One optimal global alignment (see the bold path) with score -1:

v: -TACGGGTAT-
w: GGAC--GTACG

Task 2 – Pair-wise versus profile alignment: BLAST and PSI-BLAST

An expect threshold of 1 means that we only want to see hits with an E-value less than 1. The E-value tells us how many hits of this quality we should expect by chance when using the *nr* database. E.g. the default value (10) means that 10 such matches are expected to be found merely by chance.

BLAST only returns *Populus* sequences of unknown function, so this tells us nothing about the function of our sequence.

PSI-BLAST runs in iterations. The first iteration is a regular BLAST run and therefore gives the same hits as BLAST. The hits meeting the expect threshold (i.e. 1) are separated into those that also meet the PSI-BLAST Threshold (default 0.005)(marked with a yellow label saying “New”) and those that don’t.

In the second PSI-BLAST iteration, the hits meeting the PSI-BLAST Threshold are used to build a multiple alignment and a profile. This profile is then aligned against the sequences in the *nr* database. A profile search is more sensitive than a single sequence search since we now can tell conserved positions from non-conserved positions, and thus can find significant alignments with much lower sequence similarity. Thus we get more hits, some of them indicating a possible function for our sequence.

Note that in consecutive iterations, hits meeting the PSI-BLAST Threshold are separated into those that were found in previous iterations (green dot) and newly found sequences. As we run more iterations, the profile can move away from the original sequence, especially in this case where there were very few initial hits. Hence, the new hits are taking over and by the fourth iteration the original hits do no longer appear in the list (even the hit to the sequence itself!). Clearly, too many iterations combined with a weak initial profile can move the profile so far away from the original sequence that false positive hits are retrieved.

Task 3 – Multiple alignments

From the multiple alignment it seems like our sequence have two paralogous. These are also found in the hybrid aspen (*Populus trichocarpa* x *Populus deltoides*).

Task 4 – Hidden Markov models: pFAM

pFAM finds hits to two families, although the hit is not significant according to pFAM’s definition. The function (Homeobox associated leucine zipper) is not in agreement with the PSI-BLAST hits. It might be that the weak initial profile given to PSI-BLAST (few good initial hits by BLAST) produced a false link to ATP proteins. Considering that we used the nr-database (all known sequences), this aspen protein remains something of a mystery.

Task 5 – More pair-wise sequence alignment

b) Same as in a), but \cap is a free ride.

			G	G	A	C	G	T	A	C	G	
		$S_{i,j}$	0	1	2	3	4	5	6	7	8	9
0	0	0	\cap 0	\cap 0	\cap 0	\cap 0	\cap 0	\cap 0	\cap 0	\cap 0	\cap 0	\cap 0
T	1	\cap 0	\cap 0	\cap 0	\cap 0	\cap 0	\cap 0	↘ 1	→ 0	\cap 0	\cap 0	\cap 0
A	2	\cap 0	\cap 0	\cap 0	↘ 1	→ 0	\cap 0	\cap 0	↘ 2	→ 1	→ 0	
C	3	\cap 0	\cap 0	\cap 0	↓ 0	↘ 2	→ 1	→ 0	↓ 1	↘ 3	→ 2	
G	4	\cap 0	↘ 1	↘ 1	→ 0	↓ 1	↘ 3	→ 2	→ 1	↓ 2	↘ 4	
G	5	\cap 0	↘ 1	↘ 2	→ 1	↓ 0	↘ 2	↘ 2	↘ 1	↓ 1	↘ 3	
G	6	\cap 0	↘ 1	↘ 2	↘ 1	↘ 0	↘ 1	↘ 1	↘ 1	↘ 0	↘ 2	
T	7	\cap 0	↓ 0	↓ 1	↘ 1	↘ 0	↓ 0	↘ 2	→ 1	↘ 0	↓ 1	
A	8	\cap 0	\cap 0	↓ 0	↘ 2	→ 1	→ 0	↓ 1	↘ 3	→ 2	→ 1	
T	9	\cap 0	\cap 0	\cap 0	↓ 1	↘ 1	↘ 0	↘ 1	↓ 2	↘ 2	↘ 1	

Optimal local alignment (see the bold path) with score 4:

v: GGACGT**ACG**

w: TACGGGTAT

c) Looking at the dynamic programming table in a), it is obvious that a penalty of -20 to open a gap will result in no gaps at all. Thus we will stay in the main level for the entire length of the sequence and the alignment will be:

v: TACGGGTAT
w: GGACGTACG

with a score of -7.

Task 6 – HMMs

a)

Emission probabilities	Flower	Taxicab
Pink	$\frac{3}{4}$	$\frac{1}{4}$
Yellow	$\frac{1}{2}$	$\frac{1}{2}$

Transition probabilities	Pink	Yellow
Pink to ...	$\frac{1}{2}$	$\frac{1}{2}$
Yellow to ...	$\frac{3}{4}$	$\frac{1}{4}$

b)

Note that $P(\text{sequence, path}) = P(\text{sequence} \mid \text{path}) P(\text{path})$.

Transitions used (probability in parenthesis): $P \Rightarrow Y (1/2)$, $Y \Rightarrow P (3/4)$, $P \Rightarrow P (1/2)$

Thus $P(\text{path}) = \frac{1}{2} \times [\text{Probability of transitions}] = \frac{1}{2} \times [\frac{1}{2} \times \frac{1}{2} \times \frac{3}{4}]$

$P(\text{sequence} \mid \text{path})$ is simply a matter of reading off the emission probabilities.

Emissions: $P \Rightarrow F (3/4)$ two times, $Y \Rightarrow F (1/2)$, $P \Rightarrow T (1/4)$. Thus $P(\text{sequence} \mid \text{path}) = (\frac{3}{4})^2 \times \frac{1}{2} \times \frac{1}{4}$.

Finally, $P(\{F,F,T,F\}, \{P,Y,P,P\}) = \frac{1}{2} \times [\text{Probability of transitions}] \times [\text{Probability of emissions}] = \frac{1}{2} \times [\frac{1}{2} \times \frac{1}{2} \times \frac{3}{4}] \times [(\frac{3}{4})^2 \times \frac{1}{2} \times \frac{1}{4}] = \frac{3^3}{2^{12}} = \frac{27}{4096}$ (roughly 7 permille).

c) Determine the most probable path for the sequence {Flower, Flower, Taxicab, Flower}. Assume that the path is equally likely to start in Pink and Yellow. This task can also be solved by dynamic programming and the resulting algorithm is known as the Viterbi algorithm.

The highest possible probability for all paths ending in state k with a prefix observation $X^1 \dots X^i$ (denoted s_{ki}) can be calculated as

$s_{ki} = P(X^i | state = k) \cdot \max_l [s_{l,i-1} \cdot P(state l \rightarrow state k)]$, i.e. simply replacing the sum in the forward algorithm with a max-operator. The corresponding calculations become:

State	Prefix sequence			
	F	FF	FFT	FFTF
Pink	0,375	0,140625	0,017578	0,019775
Yellow	0,25	0,09375	0,035156	0,004395

where the highest probability with a complete observation is found for state Pink. Thus the highest attainable probability for $P(\text{sequence, path})$ is approximately equal to 0,020. Since $P(\text{path} | \text{sequence}) = P(\text{sequence, path}) / P(\text{sequence})$ where $P(\text{sequence})$ is independent of the path this is the most probable path (the probability that is actually maximized by the Viterbi algorithm is the nominator of $P(\text{sequence, path}) / P(\text{sequence})$).

The following table shows the arguments to the max-operator

Transition	Prefix sequence			
	F	FF	FFT	FFTF
Pink to Pink		0,1875	0,070313	0,008789
Yellow to Pink		0,1875	0,070313	0,026367
Yellow to Yellow		0,0625	0,023438	0,008789
Pink to Yellow		0,1875	0,070313	0,008789

We have already determined that Pink is the end state of the most probable path. The table shows that the most probable way of arriving at Pink was from state Yellow. The most probable path leading to Yellow at {FFT} must have arrived from Pink. From there it could have come to Pink from either Pink or Yellow. Thus {Y, P, Y, P} and {P, P, Y, P} are two paths that both maximize the probability of observing the sequence {F, F, T, F}.

To get the lab approved, had-in your answers to Torgeir R. Hvidsten