Lab 5 – Sequence analysis

To get the lab approved, send your answers to: david.sundell@plantphys.umu.se

Task 1 – Pair-wise sequence alignment

Consider the sequences **v** = TACGGGTAT and **w** = GGACGTACG. Assume that the match premium is +1 and that the mismatch and indel penalties are -1.

Fill out the dynamic programming table for a global alignment between **v** and **w** (see lecture slides for an example). What is the score of the optimal global alignment and what alignment does this score correspond to? Hint: Draw arrows in the dynamic programming table that indicate the best choice for each entry (i.e. insertion, deletion or (mis)match). In case of ties, choose one by random. This makes it possible to backtrack one of the optimal alignments.

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

When studying non-model species such as aspen trees (Populus trichocarpa), BLAST is often used to detect homologues in model-organisms with more annotations. In plant biology, the main model organism is Arabidopsis thaliana. Consider the following aspen sequence:

>POPTR_0001s01400.1
MNPDHFDNQEAVWEWGWERCIQEPTGDTSFLDAAKATPKAQLDNMAAGTSTTSVPKTEDRRDRK
KGYDKAYRDRCREDKKRQEDELKMQAVENARLKDENESLVKEKDTILSPKLELAANVIDQLKSENHD
LKRISDHQIIRLDALTEKIASHDELKSLRDEVARLRENVNIQDPRMQEKKQLLEEHLRLANENRLLELQ
NEFYCRMIQNERHPGN

Go to http://blast.ncbi.nlm.nih.gov/Blast.cgi and click “protein blast”. Paste the sequence into the “Query Sequence” window, choose nr as your “Database” (non-redundant collection of all known sequences) and choose blastp as your “Algorithm”. Click “Algorithm parameters” at the bottom of the page. Go through the parameters by clicking “?” to get a general understanding of their effect. Set “Expect threshold” equal to 1. What does this mean?

Run BLAST. You will get a list of significant hits (E-value < 1) as well as alignments between the query and the hits. Explore the results and see if you can understand most of the statistics. What can you say about the functional role our query protein?

Open a separate tab in your browser and navigate to the “protein blast” site again. Use the same sequence and parameters, but change the algorithm to PSI-BLAST. A few new

“Algorithm parameters” appear. Check them out! Now run PSI-BLAST. This will give you hits for the first iteration. Why are they identical to the hits obtained by regular BLAST? Why are they separated into “E-value BETTER than threshold” and “E-value WORSE than threshold”?

Run another iteration of PSI-BLAST by clicking GO next to “Run PSI-Blast iteration 2 with max 500”. Now you get much more hits. Why? What can you say about the functional role of our query protein now?

Run a few more iterations. What happens with the original BLAST hits? What does this say about the possible danger of using PSI-BLAST?

Task 3 – Multiple alignments

Go back to the BLAST hits in Task 2. At the bottom of the result-page, check “Select All” and then click “Multiple alignment”. Paralogs are homologous sequences within the same genome that were separated by a gene duplication event. How many paralogs does our sequence have?

Task 4 – Hidden Markov models: pFAM

PFAM is a database of protein families containing multiple alignments and HMM representations of these. Try running our sequence from Task 2 against the pFAM HMM library: http://pfam.sanger.ac.uk/search.

Explore! What can you say about the functional role our query protein? Does it agree with BLAST/PSI-BLAST?

Optional: Task 5 – More pair-wise sequence alignment

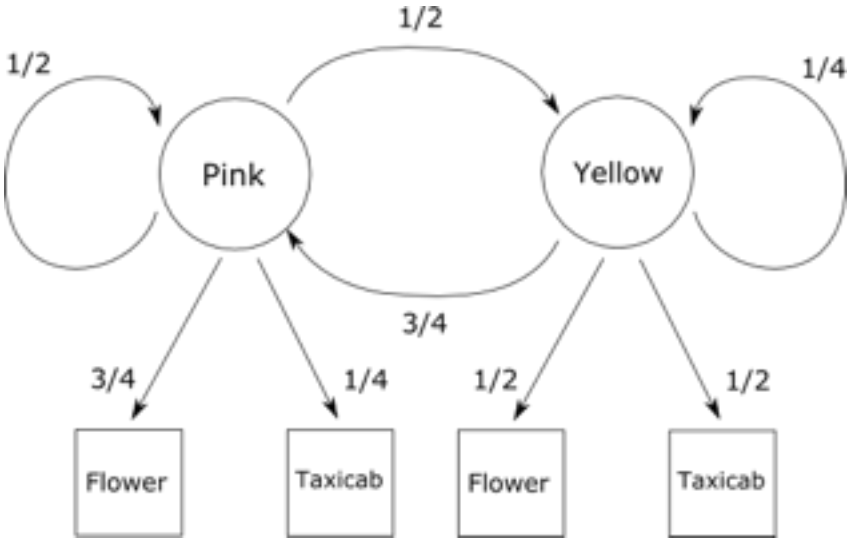
Again, consider the sequences **v** = TACGGGTAT and **w** = GGACGTACG. Assume that the match premium is +1 and that the mismatch and indel penalties are -1.

a) Fill out the dynamic programming table for a local alignment between **v** and **w**. What is the score of the optimal local alignment and what alignment does this score correspond to?

b) Suppose we use a gap penalty where it costs -20 to open a gap, and -1 to extend it. Scores of matches and mismatches are unchanged. What is the optimal global alignment in this case and what score does it achieve? Hint: You do not have to fill in the dynamic programming table to solve this task.

<u>Optional:</u> Task 6 – HMMs

Consider the HMM graphically represented below that has two hidden states, Pink and Yellow and emits two symbols, Flower and Taxicab.



a) Identify the parameters of the HMM.

b) Calculate P({Flower, Flower, Taxicab, Flower}, {Pink, Yellow, Pink, Pink}). Assume that Pink and Yellow are equally likely to be the first states.

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. I.e. use the Viterbi algorithm.