Elements of Bioinformatics, Autumn 2011, Exercise set 2

Return your answers to Esa Pitkänen (firstname.lastname@cs.helsinki.fi) on writing before the beginning of the review session, Mon 14.11 at 10am at the latest.

Exercises 2-4 can be solved by writing computer scripts to compute answers or by hand.

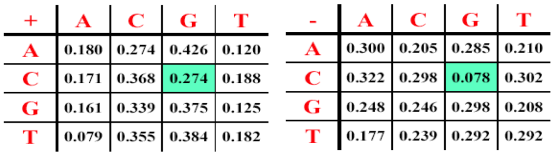
1. Fast sliding window for log-odds scores

   On slides 7-11 of Lecture 2, an approach for modelling CpG islands via Markov chains is described. Consider sliding a window of length l over the sequence to be modelled and computing for each window position the log-odds score $S_j(x) = \log \frac{P(x_j,\ldots,x_{j+l-1}|model+)}{P(x_j,\ldots,x_{j+l-1}|model-)}$.

   Derive an update formula that relieves one from computing the log-odds scores from scratch when the window position is moved from position j to $j+1$. How big is the speed-up?

2. Modelling CpG islands with Markov chains

   Below transition matrices of two Markov chains are given that are designed to recognize CpG islands (model +,left) and DNA outside the islands (model -,right).



+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

   Using the two Markov chains, compute the log-odds score $S_i(x) = \log \frac{P(x_i,\ldots,x_{i+9}|model+)}{P(x_i,\ldots,x_{i+9}|model-)}$ in a sliding window of 10 nucleotides, for the sequence given below

   ```
   TCTGTTACCCAGGCCGAGCTTC
   ```

3. Viterbi training

   Consider attending the occasionally dishonest casino with the two state HMM of slide 22 of Lecture 2.

   Consider the following sequence of rolls: 1,2,3,6,1,5,3,2,2,6,4,6,6,6,6,6,1,2,3,4,4

   Below a predicted most probable state sequence by the Viterbi algorithm for that observed sequence: F,F,F,F,F,F,F,F,F,L,L,L,L,L,L,L,F,F,F,F,F

   Use the predicted state sequence to compute the matrices A and E as well as the matrices a and e of the Viterbi training algorithm

4. Viterbi decoding

Below is a Hidden Markov Model for the occasionally dishonest casino: transition probabilities $a = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$ and emission probabilities $e = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0.17 & 0.17 & 0.17 & 0.17 & 0.17 & 0.17 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.5 \end{bmatrix}$

Using the above HMM, simulate the Viterbi decoding algorithm (slide 28, Lecture 2) using the sequence of rolls: 4 5 6 1 6 6 5

The answer should show the progression how the table v and ptr will fill, a few iterations revealing the principle are sufficient for the answer.

5. Baum-Welch algorithm

   Explain the principle of the Baum-Welch algorithm. What are the main differences to Viterbi training?

   Source: Durbin et al "Biological sequence analysis", Cambridge University Press, Chapter 3, pages 63-65; in course folder room C127, the pages 63-65 can also be found in Google books.