

## General instructions

Problems for each exercise session will be distributed approximately one week before the session. You are expected to be prepared to present your solutions in the exercise session.

## Assignments

1. Given a  $4 \times m$  PWM  $w$  and a score threshold  $R$ , the p-value is defined as the probability that a given background model produces a sequence with score greater than or equal to  $R$  when aligned with  $w$ . Let  $P(i, r)$  be the probability of getting score  $r$  for a PWM containing the first  $i$  positions of  $w$ . These values can then be evaluated using the following recursion (assuming that entries in  $w$  are small integers)

$$P(0, r) = \begin{cases} 1, & \text{if } r = 0 \\ 0, & \text{otherwise} \end{cases}$$
$$P(i, r) = \sum_{j \in \Sigma} q_j \cdot P(i-1, r - w[i, j])$$

where  $q_j$  is the probability of residue  $j$  in the background model. The p-value for  $w$  and  $R$  is then equal to  $\sum_{r \geq R} P(m, r)$ . Familiarize yourself with the method. If you are given the p-value  $\pi$ , how do you find the corresponding score threshold? (Hint: Use dynamic programming to evaluate  $P(m, r)$  for all  $r$  and use the dynamic programming array to find  $R$ .) For what values of  $i$  and  $r$  should  $P(i, r)$  be evaluated? Why do we allow only integers in  $w$ ? What is the asymptotic running time of the method?

2. Try MEME at <http://meme.sdsc.edu/meme/intro.html>. The sequences in <http://www.cs.helsinki.fi/u/lmsalmel/cmsb09/exercises/meme-data.fa> contain the CREB1 binding site. Synthesize a PWM of length 8 from this data using MEME. Compare your result with the CREB1 profile in the Jaspar database, <http://jaspar.cgb.ki.se/> (search for a profile with name CREB1).
3. Sketch an algorithm that finds from a given DNA sequence a window of width  $m$  such that this window has among all such windows the highest number of binding sites for a given set of PWMs that have score larger than  $T$ .
4. [Alon, Exercise 2.1]
5. [Alon, Exercise 2.2]
6. [Alon, Exercise 2.4]