Biological Sequence Analysis

Monday 25. February at Noon in C221

1. Sketch an algorithm that, given an IUPAC pattern (a pattern that is a sequence of normal DNA symbols and IUPAC symbols representing subsets of DNA symbols) P and a score threshold k, finds from a given DNA sequence the occurrences of P that have at most k mismatches (that is, the Hamming distance to the occurrence is at most k).

2. For the IUPAC pattern of Fig 2b [Dear p. 120], give an example of a DNA sequence that matches the pattern but does not occur in the training data given in Fig 2a.

3. Given a $4 \times m$ PWM w and a score threshold R, the corresponding p–value (= the probability that the background model produces a score equal to or greater than R) can be evaluated using the following recursion (assuming uniform model of the background and that the entries of w are small integers):

$$\begin{aligned} pval(0,r) &= \begin{cases} 1 & \text{,if } r = 0 \\ 0 & \text{,otherwise.} \end{cases} \\ pval(i,r) &= \tfrac{1}{s} \sum_{c \in \Sigma} pval(i-1, r - w[i,c]) \end{aligned}$$

   where $s = |\Sigma| =$ size of the alphabet. Familiarize yourself with this method. For what values of i and r this should be evaluated? Why only integers in w? How do you get the answer? Asymptotic running time?

4. Try MEME at `http://meme.sdsc.edu/meme/intro.html` Give to it 6 sequences from Fig 2a (Dear) and synthesize a PWM of length 12.

5. 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.