IInd. Partial exam, Information Retrieval

Curs 2009-2010, winter term

Time: 2h

Exercise 1 (1 point)

Given the pattern $TAATT$ and the text $TAx_1x_2x_3x_4x_5x_6$ where $x_i \in \{A, C, G, T\}$, we apply the BNDM algorithm and we obtain in the first iteration $D_1 = (01100)$, $D_2 = (10000)$ and $D_3 = (00000)$. Then the window is shifted obtaining the new values $D_1 = (10011)$, $D_2 = (00010)$ and $D_3 = (00000)$. What can be said about the posssible values of $x_1, \ldots, x_6$.

Exercise 2 (1.5 punts)

Given two very short strings x and y (10-20 characters). We want to find these strings but separated between 1000 and 2000 characters, that means that we are looking for paterns $x\alpha y$ where $1000 \leq |\alpha| \leq 2000$ and $\alpha, x, y \in \{A, C, G, T\}$. Design the best algorithm in two cases: exact matching (a) and approximated matching, and give the time and space complexity

Exercise 3 (1.5 points)

With the Thompson automaton look for the regular expression $A(C|G^*)$ within the sequence $AGAGA$.

Exercise 4 (1 point)

Find the factor oracle automaton for the set of patterns $AAT, ATA, TTAT$ (the automaton should be drawn with all the transitions and suffix links).

Exercise 5 (1 points)

Find for the best alignment between $ataa$ and aaa.

Exercise 6 (1.5 points)

A casino uses three tetrahedrum dice randomly choosen a each thrown. Two of them are fair and one is biased such that the value i ($i = 1, 2, 3, 4$) has the probability $i/10$.

- Which is the most probably states sequence given the results 141.

- At each thrown, which is the probability that the casino looses the bet?

Exercise 7 (0.5 points each question)

- Given a sequence of values, which is the best strategy to estimate the HMM using *esthm*?

- Is it possible to find the best alignment between 10 large sequences?

- With dotmatrix we have obtained the following picture. What can be said about the sequence/sequences?

- Advantages and disadvantages of suffix trees versus suffix arrays?