Elements of Bioinformatics, Autumn 2011, Exercise set 5

Return your answers to Esa Pitkänen (firstname.lastname@cs.helsinki.fi) on writing before the beginning of the review session, Mon 5.12 at 10am at the latest.

1. Consider the following two strings $s_1 = $ "$TCTTTTAGGA$", $s_2 = $ "$ACTTTCAGAT$".

   Compute the kernel values between the strings for the following kernels

   (a) String kernel with length-3 substrings:

   (b) Length-3 string kernel allowing one wildcard character ('?', matching any character) in the substrings

2. Draw the trie data structure that results when the matching subsequences of the above two sequences s_1 and s_2 are stored in the trie in the following two cases:

   (a) Length-3 substrings allowing no gaps

   (b) Length-3 subsequences with at most one gap

3. Gap-weighted subsequence kernels were introduced in the article H. Lodhi et al. Text Classification using String Kernels. Journal of Machine Learning Research 2 (2002), 419-444
   `http://eprints.ecs.soton.ac.uk/8968/1/String_JMLR02.pdf`

   Read the article and explain the principle behind the dynamic programming algorithm used to compute the kernel.

4. Consider the following two graphs $G_1 = (V_1, E_1)$ with $V_1 = \{v_1, v_2, v_3, v_4, v_5\}$ and $E_1 = \{(v_1, v_2), (v_1, v_4), (v_3, v_4), (v_4, v_5), (v_2, v_5))\}$, and $G_2 = (V_2, E_2)$ with $V_2 = \{v_6, v_7, v_8, v_9\}$ and $E_2 = \{(v_6, v_7), (v_7, v_8), (v_8, v_9), (v_7, v_9)\}$ with the node labels $label(v_1) = label(v_5) = label(v_6) = label(v_8) = A$ and $label(v_2) = label(v_3) = label(v_4) = label(v_7) = label(v_9) = B$.

   Compute the product graph $G_\times = G_1 \times G_2$ of these two graphs.

5. Consider the graphs G_1 and G_2 of the previous assignment. Compute the number of common length-3 walks in the two graphs.