Advanced Algorithms in Bioinformatics (P4)

# Sequence and Structure Analysis

Freie Universität Berlin, Institut für Informatik
David Weese, Sandro Andreotti
Sommersemester 2011

2. Exercise sheet, 20. April 2011
Discussion: 27. April 2011

*Exercise 1.*

Using Ukkonens algorithm for k-differences matching, find all occurrences of the pattern $P = \texttt{tcaa}$ in the text $T = \texttt{atcatcaatc}$ with up to $k = 2$ differences. Show the dynamic programming matrix, the value of $lact$ for each column, do not compute unnecessary cells, etc.. For one column of your choice, keep track of the auxiliary variables $Cn$ and $Cp$ as well as the whole column vector $C$ (so as to understand their meaning).

*Exercise 2.*

Prove the correctness of the following observations mentioned in the lecture. $C$ is a dynamic programming matrix computed using the edit distance.

$$\begin{aligned} \text{horizontal adjacency property} \quad & \Delta h_{i,j} = C_{i,j} - C_{i,j-1} \in \{-1, 0, +1\} \\ \text{vertical adjacency property} \quad & \Delta v_{i,j} = C_{i,j} - C_{i-1,j} \in \{-1, 0, +1\} \\ \text{diagonal property} \quad & \Delta d_{i,j} = C_{i,j} - C_{i-1,j-1} \in \{0, +1\} \end{aligned}$$

(Hint: induction – on what?).

*Exercise 3.*

Conclude from Exercise 2 (you may use it even if you have not done the proofs) that the value of $lact$ (in Ukkonen's algorithm for string matching with $k$ differences) can increase in one iteration by at most one.

*Exercise 4.*

This time use Myers' bit-vector algorithm for pattern and text in Exercise 1.
Describe how to extend Myers' bit vector algorithm for approximate string matching with patterns larger than the word length. Describe you approach as pseudocode.