Advanced Algorithms in Bioinformatics (P4)

Sequence and Structure Analysis

Freie Universität Berlin, Institut für Informatik
Prof. Dr. Knut Reinert, Sandro Andreotti
Sommersemester 2009

2. + 3. Exercise sheet, 29. April 2009
Discussion: 8. May 2009

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.

This time use Myers' bit-vector algorithm for pattern and text in Exercise 1.

Exercise 3.

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

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

(Hint: induction – on what?).

Exercise 4.

Conclude from Exercise 3 (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 5.

The following lemma is central to the PEX algorithm:

Lemma 1. *Let Occ match P with k errors, $P = p^1, \ldots, p^j$ be a concatenation of subpatterns, and $a_1, \ldots, a_j$ be nonnegative integers such that $A = \sum_{i=1}^{j} a_i$. Then, for some $i \in 1, \ldots, j$, Occ includes a substring that matches p^i with $\lfloor a_i k/A \rfloor$ errors.*

1. Following this Lemma show by formal substitution:
   (a) Let Occ match P with k errors and $P = p^1, \ldots, p^{k+1}$ be a concatenation of subpatterns. Then at least one of the p^i matches Occ exactly, for some $i \in 1, \ldots, k+1$.
   (b) Let Occ match P with $2k+1$ errors and $P = p^1, \ldots, p^{k+1}$ be a concatenation of subpatterns. Then at least one of the p^i matches Occ with at most one error, for some $i \in 1, \ldots, k+1$.
2. Prove Lemma 1.

Exercise 6.

Find the pattern $P =$ `filter` in the text $T =$ `pex_hierarchical_verification_filter` with at most $k = 2$ errors. Compare the verification costs of non-hierarchical filtering directly following Lemma 1 (split pattern into $k+1$ subpatterns and search for perfect matches) and the PEX algorithm.

Exercise 7.

The following lemma is central to the (ungapped) Quasar algorithm. Prove it.

Lemma 2. *Let P and S be strings of length w with at most k differences. Then P and S share at least $w + 1 - (k+1)q$ common q-grams.*