Advanced Algorithms in Bioinformatics (P4)

Sequence and Structure Analysis

Freie Universität Berlin, Institut für Informatik
Knut Reinert, Sandro Andreotti
Sommersemester 2012

4. Exercise sheet, 8. May 2012
Discussion: 23. May 2012

Exercise 1.

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 2.

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 3.

The following (q-gram) 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.*

Exercise 4.

Find a gapped shape of size at least fore and value of w such that the generalization of the q-gram Lemma for gapped shapes does not yield a tight threshold (>= 0)