Universität Bielefeld
Technische Fakultät

AG Genominformatik
Prof. Dr. Jens Stoye

Algorithms in Genome Research
Winter 2021/2022

Exercises

Number 5, Discussion: 2021 December 17

1. Given the two sequences $x =$ `AATTGCC` and $y =$ `ATGTGAATG`, consider the following score function: match = 4, mismatch = -3, indel = -2.
   (a) Compute all optimal global alignments of x and y.
   (b) Compute all optimal free-end gap alignments of x and y.
   (c) Compute all optimal local alignments of x and y.
2. Affine gap costs for a gap of length ℓ are defined as $g(\ell) = d + e(\ell - 1)$ where d is the *gap initiation cost* and e is the *gap extension cost*.
   (a) Why should d not be chosen lower than e?
   (b) Show that affine gap costs are *subadditive*, i.e. $g(\ell_1 + \ell_2) \leq g(\ell_1) + g(\ell_2)$.
   (c) Compute an optimal global alignment with affine gap costs of the sequences $x =$ `ATCCTAG` and $y =$ `ATTGCCCT`, using the following scoring scheme: match = 3, mismatch = -2, gap open $d = 3$, gap extension $e = 1$.
3. Let $p =$ `TCAG` be a pattern string and $t =$ `AACGTCAGTCGAGTG` be a text string.
   (a) Find all positions in t where an approximate occurence of p with up to $k = 1$ errors ends.
   (b) For each of these positions, give one corresponding alignment.
   (c) Why is it useful in practice not to return *all* end positions? In the above example, which positions should be reported?
4. (from exercise sheet 3)
   In an assembly strategy that is based on long reads only, overlapping long reads have to be detected. What are the challenges, and how could a dynamic programming algorithm be designed that solves this problem?