Universität Bielefeld
Technische Fakultät

Prof. Veli Mäkinen, PhD
Linda Sundermann

Algorithms in Genome Research
Winter 2012/2013
Exercises

Number 5, Discussion: 23. November 2012

1. Construct the Burrows-Wheeler transform of `ACATGATCTGCATT` and simulate backward search on it with pattern `CAT`.
2. Construct the Burrows-Wheeler transform of `ACATGATCTGCATT` and simulate 1-mismatch backward backtracking search on it with pattern `CAT`.
3. How long should be the read length m so that there are no occurences of a read expected in a random sequence of length n allowing k mismatches?
4. Some mate pair sequencing techniques work by having an adapter where the two tails of a long DNA fragment bind to, forming a circle. This circle is cut in one random place and then X nucleotides apart from it again, forming one long fragment and another shorter one (assuming X much smaller than circle length). The fragments containing the adapter are fished out from the pool (together with some background noise). Then these adapter containing fragments are sequenced from both ends to form the mate pair. Because of the random process of cutting, some of the mate pair reads may overlap the adapter. Such overlaps should be cut before using the reads any further.
   a) Give an algorithm to cut the adapter from the reads. Take into account that short overlaps may appear by chance and that the read positions have the associated quality values denoting the measurement error probability.
   b) How can you use the information about how many reads overlap the adapter to estimate the quality of fishing?