Universität Bielefeld
Technische Fakultät

Prof. Veli Mäkinen, PhD
Linda Sundermann

Algorithms in Genome Research
Winter 2012/2013

Exercises

Number 7, Discussion: 7. December 2012

1. Consider different large scale variations in genomes, like gene dublication, copy number variation, inversions, translocations, etc. How can they be identified using read mapping? Is there an advantage of using paired end reads?

2. Consider an array $A[1, n]$ and a query $A.\text{RangeCount}(l, r, i, j) = |\{k \mid i \leq A[k] \leq j, l \leq k \leq r\}|$.

   a) Show that the balanced wavelet tree of Sect. 3.4 build on A can support the RangeCount query in $O(\log n)$ time.

   b) How can this query be exloited in RNA-sequencing read alignment and in paired end read alignment? *Hint. Use the structure on suffix array of the reference genome.*

3. Consider the overlap computation with stacks and suffix tree as described in Sect. 4.2.3. Assume you have a compressed suffix tree representing the concatenation of reads in small space, and it supports depth-first travelsal, retrieval of string depth, etc. common suffix tree operations efficiently. How much extra space do you need in the worst case for the stacks, doubly-linked lists, etc. structures to implement the overlap computation on top of the given compressed suffix tree? Do you find any compression methods to improve the extra space requirement? *Hint. There are dynamic bit-vectors taking $O(N)$ bits space to represent a bit-vector of length N allowing logarithmic time inserts and deletions. One can imagine representing stacks (whose values are increasing) with them, and concatenation of stacks with another boundary vector.*

4. Modify Algorithm maximalRepeatsTwoWayBWT() as suggested at page 57 / Sect. 4.2.2 to solve maximal exact matches problem with two strings. Simulate the algorithm with an example of your choice (you can restrict to showing one step of the algorithm as done in the lecture).