Universität Bielefeld
Technische Fakultät

AG Genominformatik
Prof. Dr. Jens Stoye
Dr. Marília Dias Vieira Braga

Algorithms in Genome Research

Winter 2023/2024

Exercises

Number 4, Discussion: 2023 December 01

1. For identifying the overlap between two reads, one approach consists in finding the best prefix-suffix alignment between them. Write down the details of the recursion of a dynamic programming algorithm for input sequences s_1 and s_2, given a matrix σ with the scores of matches/mismatches for all pairs of symbols $x, y \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and a constant gap cost c.

2. Discuss the reasons why the traditional overlap-layout-consensus (OLC) assemblers fail to assemble short-read data.

3. Questions around assembly with high throughput sequencing – short and long reads:

   (a) The basic data structure used for short-read sequence assembly is the de-Bruijn subgraph, defined for a given set of reads and a given dimension k. While the de-Bruijn subgraph is conceptually easy, there are several challenges when you want to implement it in practice – name a few.

   (b) What happens to the de-Bruijn subgraph when the dimension k is increased/decreased?

   (c) A DNA sequence is a *palindrome* when it is identical to its reverse complement. What restriction can be applied to k for avoiding the occurrence of palindromic k-mers?

   (d) What can be done with long reads that cannot be done with mate pairs?

4. Draw the 4-dimensional de-Bruijn subgraph (i.e. where vertices correspond to 4-grams) for the following set of reads.

   `AAATG`, `AATGA`, `AATGAC`, `AATGC`, `ACCAG`, `ACCAGA`, `ACCTG`, `ACGTT`, `AGACG`, `AGACGG`, `ATAAT`, `ATAATG`, `ATAATGC`, `ATGAC`, `ATGCA`, `ATGCAC`, `CACGG`, `CAGAC`, `CCAGA`, `CGTTA`, `CTGACGT`, `GACCA`, `GACCAGA`, `GACGTT`, `GCACG`, `GCACGG`, `GTTAAT`, `GTTAATG`, `TAATG`, `TAATGA`, `TACTA`, `TGACC`, `TGCAC`, `TTAAT`.

   Can you assemble the data set into a single contig? (There may be some "sequencing errors" that need to be corrected.)

5. What are the main differences (input, major steps, advantages/disadvantages) between "de-novo" genome assembly (via OLC) and comparative assembly?

(continue in the next page)

6. Let the following DNA sequence be a *reference genome*:

   `AATGAGGTCATCCTTGCTGGACTCTAGCAC`

   The following three sets of reads (a), (b) and (c) originate from three distinct *target genomes* that are closely related to the reference. Each target genome differs from the reference by a single structural variation (rearrangement or indel).

   Reconstruct the three target genomes by mapping the reads to the reference and identify the rearrangements. (Assume that there are no sequencing errors and recall that a read may come from any of the two complementary DNA strands.)

(a)			(b)			(c)		
	1	ACTCTAGCAC		1	AATGACAAGG		1	AATGAGGTCA
	2	AGTCCTGTACAG		2	ACCCTGGACTCT		2	AGGTCATCGAC
	3	CCTTGCTGTA		3	GGATGACCCTG		3	AGTCGATGAC
	4	GCTGTACAGGAC		4	GTCATCCTTG		4	CATCGACTCT
	5	GGTCATCCTT		5	GTGCTAGAGT		5	CTAGAGTCGAT
	6	TGACCTCATT		6	TCCAGGGTCA		6	GTGCTAGAGT