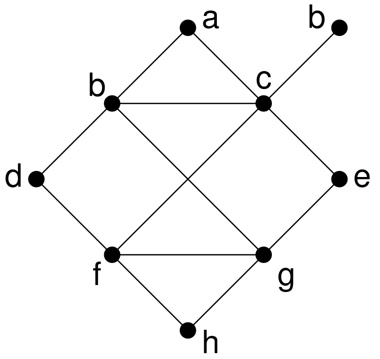
Bioinformatik für Biologen SS 2019

Complementary Homework 12

This homework will not be collected or graded, but its contents and the papers to read may be relevant in an upcoming lecture, test or assignment.

1. Analyze the following Graph:



- does an Euler Path exist for this graph?
- does an Euler Cycle exist for this graph?
- write down the sequence of nodes for an Euler Path and/or Euler Cycle, if it exists, respectively.
- write down degrees of the nodes.

Add-on: Does an Euler Path and/or Euler Cycle exist for the Königsberg-Brigde-Problem (i.e. the original map of the city)?

2. Two research groups study flies with a genome size of about 150mio nucleotides and both had the genome sequenced.
One group got sequencing data from a PacBio machine that produced 60000 reads with average length 10000nt.
The other group obtained sequencing data from an Illumina HiSeq machine. This sequencer sequenced 500mio fragments and for each fragment it obtained two paired-end reads of length 100nt each.

- compute the coverage for each of the sequencing data.
- discuss briefly which are the advantages and disadvantages of each dataset for the assembly process.
- which obstacles can be overcome if a reference genome exists for the assembly of the single-end data?

3. You have sequenced 16 times the read ACGTATCGA and once ACGTTTCGA.

- generate a de Bruijn Graph for the reads with k=4.
- what structure does that graph contain?
- what would happen to that structure during the subsequent steps in the assembly, and why?
- how do you distinguish 'real' paths in a de Bruijn Graph from those caused by sequencing errors?

4. You are analyzing a hypothetical genome with 7 regions:
region1-regionR-region2-regionR-region3-regionR-region4
regions with the same name are exact repeats.
- draw the sketch of an condensed de-Bruijn-Graph which would result from this genome.
- in how many separate contigs would one reconstruct?
- how many paths through the graph exist to resolve the genomic structure?
- are there possible ways to figure out which path is the correct one?
- is there a way to infer how often a repeat existed in the originally sequenced template?

5. You have done a de-novo assembly and you have got 14 contigs with the following lengths:
100, 1000, 200, 2000, 2500, 300, 3000, 500, 4000, 600, 700, 800, 8000, 900
What is the N50 value?
What is the difference to an NG50 value?

6. You want to use seeds to index a genome to search for candidate hit regions to map reads.
- Do explain the difference between contiguous and spaced seeds.
- What is the potential problem with contiguous seeds?
- Why is using sets of spaced seeds typically more sensitive than using just one seed (spaced or contiguous)?

7. Analyze the 'genome' ACACG using a BWT.
- generate the BWT from the sequence.
- state the lexicographical order you used.
- decode the original sequence from the BWT using the first-last-columns approach.
- search for the strings AC and CC in it using the BWT-approach.
- search for the strings CC using the BWT-approach allowing 1 mismatch.

8. Draw a simple eukaryotic protein-coding gene model containing (a) introns, (b) exons, (c) promoter, (d) UTRs, (e) poly-A signal, (f) translation start site, (g) transcription start site and (h) stop codon. Also mark the (i) coding region(s) of the gene.
Furthermore, draw the according mature mRNA and mark the elements above if they exist in the mRNA.

9. Explain the difference between extrinsic and intrinsic methods for gene prediction.
- Think of advantages of each of the two approaches.
- What is the advantage of using ESTs or cDNA with an extrinsic method to predict eukaryotic genes, compared to the use of a protein database?
- Name genomic features that can be used by an intrinsic approach to predict eukaryotic genes.
- Why is it hard (if not impossible) to build a method perfectly predicting genes in all species.