Literature assignment 2 **Due: Nov. 3rd, 2016 at 4:00pm**

Your name:

Article:

- Phillip E C Compeau, Pavel A. Pevzner, Glenn Tesler. *How to apply de Bruijn graphs to genome assembly*. Nature Biotechnology 29, 987-991(2011), doi:10.1038/nbt.2023.

Read this article including the supplementary text and answer the following questions. You may read additional materials, if you wish. If you do, you must cite your sources. *You may not quote verbatim without attribution.*

I. Hamiltonian and Eulerian paths are two graph theoretic approaches that have been used in short-read sequence assembly. Here you are asked to apply them to small examples in a variety of situations.

1. Let

CAAAAGTGCACCGGTTC

be a genome sequence consisting of a single *linear* chromosome. Imagine that you sequence this genome using an error-free sequencing machine that generates reads of length 5. Your sequencing run results in full coverage; that is, you obtain at least one read that starts at each position in the genome. Show the reads you obtain from this sequencing run.

2. If all reads are fragmented into 3-mers, how many *unique* 3-mers will result? Show them. Are there any 3-mers in the genome that are not included in the set you obtained from the reads?

3. Construct the directed graph (digraph) for the set of unique 3-mers resulting from these reads, based on overlapping suffixes and prefixes, as shown in Fig. 3(b) of Compeau *et al.* (2011)

How many Hamiltonian paths are there in this graph? Note that these path(s) will not necessarily be cycles.

Show

- the digraph,
- alll Hamiltonian path(s), and
- the assembly corresponding to each path.

4. Was the correct genome sequence represented among the assemblies that you found? If so, was it unique? If you did not find a unique, correct assembly, what aspect of the sequence or the graph might be responsible for this?

5. Construct a de Bruijn graph for the same set of 3-mers, as shown in Fig. 3(d) of Compeau *et al*. (2011). Is this de Bruijn graph balanced? If not, which node(s) cause the imbalance?

   How many Eulerian paths are there in this graph? Again, the path(s) will not necessarily be cycles.

   Show
   - the de Bruijn graph,
   - the Eulerian path(s), and
   - the assembly corresponding to each path.

6. Was the correct genome sequence represented among the assemblies that you found? If so, was it unique? If you did not find a unique, correct assembly, what went wrong?

II. The impact of increasing *k* on the Hamilton and Eulerian paths: Does increasing *k* from 3 to 4 lead to a better assembly?

1. You attempt to assemble the same set of reads using 4-mers, instead of 3-mers. If all the reads from Question I are fragmented into 4-mers, how many *unique* 4-mers will result? Show them. Are there any 4-mers in the genome that are not included in the set you obtained from the reads?

2. Construct the directed graph (digraph) for the set of unique 4-mers resulting from these reads, based on overlapping suffixes and prefixes, as shown in Fig. 3(b) of Compeau *et al.* (2011)

   How many Hamiltonian paths are there in this graph? Note that these path(s) will not necessarily be cycles.

   Show

   - the digraph,
   - the Hamiltonian path(s), and
   - the assembly corresponding to each path.

3. Was the correct genome sequence represented among the assemblies that you found? If so, was it unique? If you did not find a unique, correct assembly, what aspect of the sequence or the graph might be responsible for this?

4. Construct a de Bruijn graph for the same set of 4-mers, as shown in Fig. 3(d) of Compeau *et al*. (2011). Is this de Bruijn graph balanced? If not, which node(s) cause the imbalance?

   How many Eulerian paths are there in this graph? Again, the path(s) will not necessarily be cycles.

   Show

   - the de Bruijn graph,
   - the Eulerian path(s), and
   - the assembly corresponding to each path.

5. Was the correct genome sequence represented among the assemblies that you found? If so, was it unique? If you did not find a unique, correct assembly, what went wrong?

III. Assembly with missing data

1. Suppose that you sequence the same genome

   CAAAAGTGCACCGGTTC

   using a lossy sequencing machine. The reads this machine produces are error-free, but some reads are lost. As before, the read length of this machine is 5.

   In this particular sequencing run, you only obtain reads starting at the following sites in the sequence: 1, 2, 3, 4, 5, 7, 8, 9, 10, and 13. What are the reads you obtain from this run?

2. If these reads are fragmented into 4-mers, how many *unique* 4-mers will result? Show them. Are there any 4-mers in the genome that are not included in the set you obtained from the reads?

3. Construct the directed graph (digraph) for the set of unique 3-mers resulting from these reads, based on overlapping suffixes and prefixes, as shown in Fig. 3(b) of Compeau *et al.* (2011).

   How many Hamiltonian paths are there in this graph?

   Show
   - the digraph,
   - the Hamiltonian path(s), and
   - the assembly corresponding to each path.

4. Was the correct genome sequence represented among the assemblies that you found? If so, was it unique? If you did not find a unique, correct assembly, what aspect of the sequence or the graph might be responsible for this?

5. Construct a de Bruijn graph for the same set of 3-mers, as shown in Fig. 3(d) of Compeau *et al*. (2011). Is this de Bruijn graph balanced? If not, which node(s) cause the imbalance?

   How many Eulerian paths are there in this graph? Again, the path(s) will not necessarily be cycles.

   Show
   - the de Bruijn graph,
   - the Eulerian path(s), and
   - the assembly corresponding to each path.

6. Was the correct genome sequence represented among the assemblies that you found? If so, was it unique? If you did not find a unique, correct assembly, what went wrong?

IV. Assembly with dispersed repeats

1. Suppose you sequence a *circular* genome

GCAGTCCAGGC

on an error-free sequencing machine. What are the reads you obtain from this run?

2. If all reads are fragmented into 3-mers, how many unique 3-mers will result? Is this the complete set of 3-mers in the genome?

3. Construct the directed graph (digraph) for the set of unique 3-mers resulting from these reads, based on overlapping suffixes and prefixes, as shown in Fig. 3(b) of Compeau *et al.* (2011).

   How many Hamiltonian paths are there in this graph?

   Show
   - the digraph,
   - the Hamiltonian path(s), and
   - the assembly corresponding to each path.

4. Was the correct genome sequence represented among the assemblies that you found? If so, was it unique? If you did not find a unique, correct assembly, what aspect of the sequence or the graph might be responsible for this?

5. Construct a de Bruijn graph for the same set of 3-mers, as shown in Fig. 3(d) of Compeau *et al*. (2011). Is this de Bruijn graph balanced? If not, which node(s) cause the imbalance?

   How many Eulerian paths are there in this graph?

   Show
   - the de Bruijn graph,
   - the Eulerian path(s), and
   - the assembly corresponding to each path.

6. Was the correct genome sequence represented among the assemblies that you found? If so, was it unique? If you did not find a unique, correct assembly, what went wrong?

7. Using the strategy proposed in Box 2 of Compeau *et al.* (2011), ("Handling DNA repeats"), construct the modified de Bruijn graph for the unique 3-mers of this sequence. Is this de Bruijn graph balanced? If not, which node(s) cause the imbalance?

   How many Eulerian paths are there in this graph?

   Show
   - the graph (with edges labeled with the 3-mers),
   - the Eulerian path, and
   - the corresponding assembly, assuming that your sequence is linear, not circular.

8. Was the correct genome sequence represented among the assemblies that you found? If so, was it unique? If you did not find a unique, correct assembly, what went wrong?