Genome Assembly Definitions

read

k-mer

read length

contig

assembly

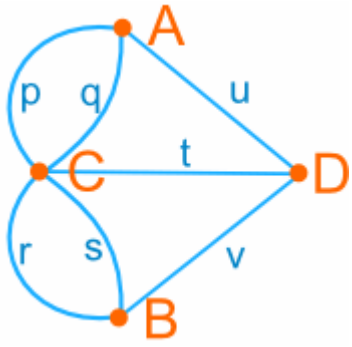
coverage

Notes

In an overlap graph, nodes are _____ and edges are _____.

If the sequence is a circle, we find a _____. Otherwise, we find a _____.

The assembly in an overlap graph is a _____ cycle/path.



This cycle/path exists if and only if:

There is/is not (circle one) a cycle crossing over all seven Bridges of Königsberg because:

In a de Bruijn graph, nodes are _____ and edges are _____.

The assembly in a de Bruijn graph is a _____ cycle/path.

This path is easier/harder (circle one) to find than finding a path in the overlap graph.

To reconstruct a sequence with maximum repeat length *n*, *k*-mers must be _____ long.

Assembly Lessons from Dr. Seuss

GOAL

Reconstruct the text by lining up the fragments of text in the proper order.

RULES

1. Each word in the text may appear more than once in the fragments (i.e. fragments may overlap).

2. If you have a reference, it may not match the fragments exactly.

3. Fragments with a number in parentheses mean that there are exactly (n) words in between the phrase on the left and the phrase on the right.

4. If you have heard the text elsewhere, it is fair to use that knowledge. Mention this in your answers.

5. However, do not look up the text online.

QUESTIONS FOR EACH TEXT

Describe the fragments.

Can you definitively reconstruct the text? Is the task easy or hard?

Is the problem easy/hard due to the text, the fragments, or both; and in what ways?

BONUS

Describe your method for solving the problems. This does NOT need to be a perfectly specific algorithm.

Did you have different methods for different texts?

Constructing Assembly Graphs

For each set of k-mers:

1. Construct the de Bruijn graph
   For set 3 ONLY, condense the de Bruijn graph.
2. Find the Hamiltonian cycle/path (or one of them) and write the assembly
3. Construct the overlap graph
4. Find the Eulerian cycle/path (or one of them) and write the assembly

1	2	3
CAA	**GGCGTGC**	**ACT**
AAT	**TGCAATG**	**GAC**
ATG	**CAATGGC**	**ACT**
GCA	**ATGGCGT**	**CCG**
TGC	**CGTGCAA**	**CGA**
GCG		**CTC**
TGG		**TCC**
CGT		**ACT**
GTG		**CTT**
GGC		**GGG**
		AAG
		AGA
		GAC
		CTG
		GAC
		GGA
		TGG
		TTT