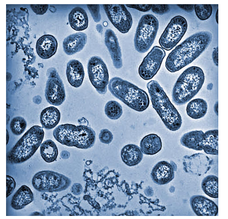
Reconstructing the *Salmonella bongori* Genome from Paired Reads

Jen Johnson

Middlebury College, CSCI 321 Spring 17



Goal

The goal was to construct contigs of the *Salmonella bongori* genome from error-free read pairs using a paired de Bruijn graph and compare the number and length of the contigs with those produced from single reads.

Background

De Bruijn graphs are a method of reconstructing a genome from short reads of DNA. Contigs are nonbranching paths in a de Bruijn graph that represent a region of certainty in the genome. The accuracy of a de Bruijn graph is measured by the length and number of contigs that can be formed from it. The most accurate graphs have a high number of contigs of long length.

Paired reads are 2 k-mers separated by a unsequenced region of length d. Medvedev *et al*. (2011) claim that paired De Bruijn graphs facilitate genome reconstruction. This is because there is a lower chance of k-mer overlap because both reads in the pair must match by prefixes and suffixes. Therefore, the paired graphs are less tangled and have a higher number of longer contigs.

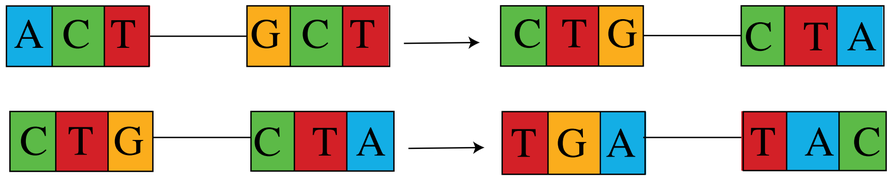
Because of these advantages, paired-read genome reconstructions are considered to be higher quality and have a higher measure of confidence than single reads reconstructions. Therefore, single read reconstructions and paired-read reconstructions are compared to evaluate this claim.

Methods

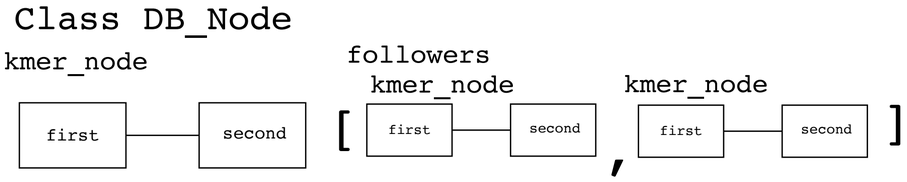
The approach to construct contigs from paired reads is as follows:

1. Turn the genome string into paired reads of length 2k+d and store them as kmer_nodes.

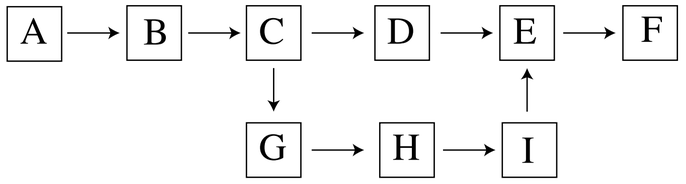
2. Form an overlap graph using the prefixes and suffixes of both reads in each paired read.



3. Form a paired De Bruijn graph in the form of an adjacency matrix by storing kmers as an array of db_nodes and gluing identically-labeled nodes.



4. Generate contigs from the De Bruijn Graph.



Contigs:
ABC
CDE
CGHIE
EF

Results

Salmonella bongori is a rod-shaped bacteria that causes the disease salmonellosis. It was chosen because prokaryotic genomes are shorter than eukaryotic genomes and therefore more feasible for the scale of this project.

It was found that the number of contigs generated using paired reads was less than the number generated using single reads. This did not support Medvedev's observation. It was found that the average length of paired read contigs was greater than that of single read contigs for all length of genomes tested.

Length of Genome in Nucleotides	Average Length of Contigs in Nucleotides	
	Single Reads	Paired Reads
17	3.7	4.7
350	26.9	32
700	45.3	382
1400	51	1082
3500	35.9	3182

Table 1: Average Length of Contigs Produced using Single and Paired Reads. The genomes were fragments of the *Salmonella bongori* genome where k = 10 and d = 300.

This could explain the observation for the number of contigs. If the contigs found are longer, there does not need to be as many contigs to achieve the same level of coverage. The observation for the length of contigs supports Medvedev's observation that paired de Bruijn graphs produce longer contigs than single read de Bruijn graphs, and therefore paired de Bruijn graphs are more valuable for genome reconstruction.

References

https://github.com/jenjohnson7/CSCI321FinalProject

Illumina.com. "Paired-End Sequencing." https://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html. Accessed 12 May 2017.

Medvedev, P., Pham., S., Chaisson, M., Tesler, G., and Pevzner, P. "Paired De Bruijn Graphs: A Novel Approach for Incorporating Mate Pair Information into Genome Assemblies." Journal of Computational Biology 18, no 11 (2011):1625-1534. Accessed 10 May 2017. DOI: 10.1089/cmb.2011.0151.

Vumicro.com. "Salmonella bongori."Accessed 12 May 2017. http://www.vumicro.com/vumie/help/VUMICRO/Salmonella_bongori.htm