Reconstructing the Carsonella ruddii genome using a paired de Bruijn graph.

Julia Fairbank, Mia Tarantola and Caroline Cutter
CS321 Bioinformatics, Professor Linderman

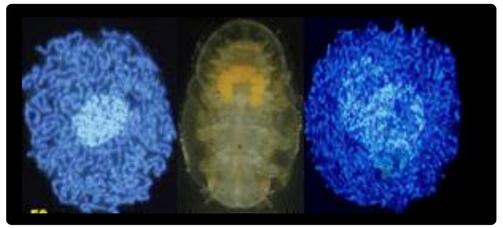
Goal

Reconstruct the genome of *Candidatus Carsonella ruddii* using a paired de Brujin graph from error-free coverage and error-free read pairs.

Background

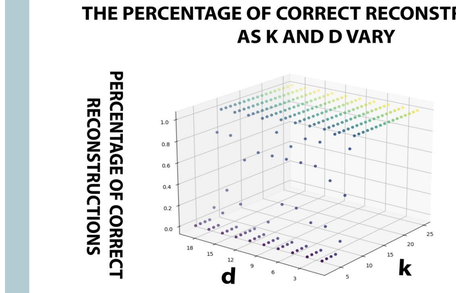
The bacteria *Ca. Carsonella ruddii* contains one of the smallest genomes identified, with only ~160,000 base pairs and 182 genes, surviving by making use of the host's genome.

The genome is so small that many biologists claim that it lost its "bacterial" identity and transitioned to an organelle, a recurring transition in evolution history.

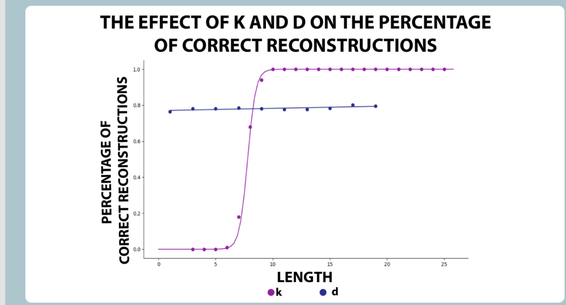


Methodology & Reconstruction

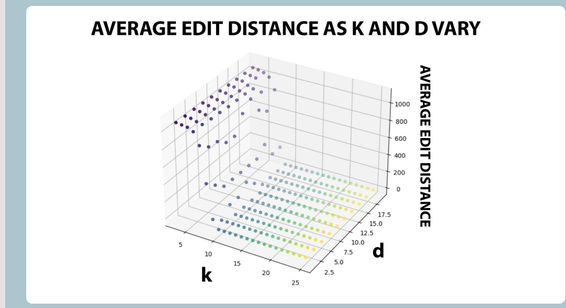
1. Randomly select 10 read of length 1,000 from the 166,000 long genome
2. Generate the (k, d)-mer composition
3. Generate the paired De Bruijn Graph
4. Create a new edge from the end node to start node and create Eulerian Cycle
5. Break the cycle and generate Eulerian Path
6. Reconstruct the sequence
7. Repeat process with varying k and d values



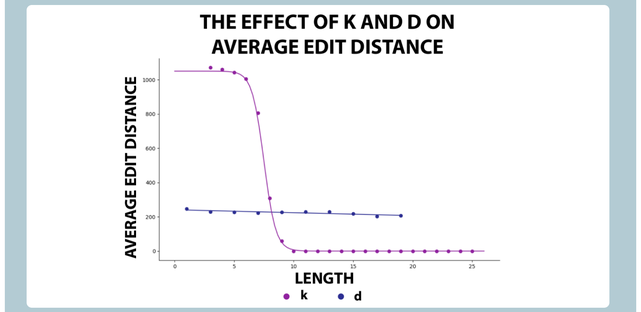
ANALYSIS: The percentage of correct reconstructed strings (PCRS) is largest for (k,d)mers with high k values, and smallest for (k,d)mers with low k values. So, k is proportionate to PCRS.



ANALYSIS: PCRS remains constant as d varies. PCRS remains constant as k increases from 1 to ~8, until a rapid increase, then plateaus. PCRS is dependent on k, not d. This suggests that the minimum k for correct reconstruction is ~8.



ANALYSIS: The avg. edit distance (AED) for (k,d)-mer pairs with a large k value is the smallest, and largest for the pairs with small k values, suggesting that k is roughly inversely proportionate to AED.



ANALYSIS: The AED across all values of d remains constant. While the AED is constant with a rapid decline then constant again. This suggests the optimal value of k is >=~8. AED only depends on k.

Conclusions

- The accuracy of this methodology is dominated by k rather than d
- K is proportionate to PCRS, but inversely proportionate to AED.
- Neither PCRS or AED are dependent on d, as the largest source of error is repeated k-mers, which is also not dependent on d
- After determining the longest repeated kmer is length 8, the results concur that the minimum k for a high percentage of correctness is 8/9.

References

National Library of Medicine, "Candidatus Carsonella ruddii HC isolate Thao2000, complete sequence"
"Tutorials — Matplotlib 3.5.2 Documentation."