R. Cox, K. Long, H. Cole
394P Bioinformatics Project Proposal
3.25.2019

De Bruijn Assembly of Proteomes

Modern short-read assembly algorithms commonly use de Bruijn graphs (DBG) for *de novo* assembly of genomes. During traditional assembly of a genomic DBG, reads are broken into smaller fragments of a specified size to generate *k*-mer prefixes and suffixes. Prefixes and suffixes (i.e., “(*k*-1)mers”) are used as nodes and edges represent *k*-mers having a specific prefix and suffix. Several desirable properties are associated with DBGs, but the primary ones we are interested in are: (1) finding a cycle visiting all edges of a graph (i.e., a Eulerian path) is a trivial computational exercise compared to finding a cycle visiting all nodes of a graph (i.e., a Hamiltonian path), thus significantly reducing potential run time and (2) DBG construction can “keep track” of alternative paths resulting from point mutations, sequencing errors, or multiplicities (repeats) as “bulges” in the path.[1,2]

With these considerations in mind, we propose using DBG assembly of protein sequences for domain homology analysis. Since there are three of us in the group, we have generated three idea branches to pursue:

1. [Rachael Cox] Visualization of domain convergence/divergence for all orthogroups annotated in the OMA database using a large graph layout (LGL), and exploration of any interesting evolutionary patterns.[3]
2. [Kimber Long] The de Bruijn graph for this project was created in order to look at homology; however, protein homologs can have sequence similarity as low as 15%. In order to look at true divergence, a BLOSUM matrix should be incorporated, giving variations in amino acid sequence a “weight.” Based on this weight, k-mers with chemically similar amino acids may be determined to be similar and the edges will converge. Alternatively, if k-mers are determined to have sufficiently different amino acids then the edges will diverge. Computational precedent can be found in reference 2.
3. [Hannah Cole] Since its invention in the 1950's, humans have created billions of tons of plastic waste. These plastics are not biodegradable and will remain in the environment for hundreds of years. One promising method to manage this waste involves the engineering of plastic degrading enzymes. Scientists first discovered PETase activity in the bacterium Ideonella sakaiensis, but have subsequently found similar enzymes in other species of bacteria, fungi, and even wax worm caterpillars. Using our de Brujin graph workflow, I want to investigate whether there are conserved domains in the polymer-degrading enzymes found in all these species and visualize common motifs in Pymol to investigate their potential as targets for protein engineering.

References

1. Compeau, P. E. C.; Pevzner, P. A.; Tesler, G. Why Are de Bruijn Graphs Useful for Genome Assembly? *Nat Biotechnol* **2011**, *29* (11), 987–991. https://doi.org/10.1038/nbt.2023.
2. Patwardhan R, Tang H, Kim S, Dalkilic M. An Approximate de Bruijn Graph Approach to Multiple Local Alignment and Motif Discovery in Protein Sequences. In: Dalkilic MM, Kim S, Yang J, eds.

Data Mining and Bioinformatics. Vol 4316. Berlin, Heidelberg: Springer Berlin Heidelberg; **2006**:158-169. doi:10.1007/11960669_14

3. Adai AT, Date SV, Wieland S, Marcotte EM. LGL: Creating a Map of Protein Function with an Algorithm for Visualizing Very Large Biological Networks. *Journal of Molecular Biology*. **2004**;340(1):179-190. doi:10.1016/j.jmb.2004.04.047