Final Project Proposal

Michael Hoffert and Sophie Gunn

Layla Oesper
Computational Biology

February 18, 2018

1 Overview of Binning Tools

For the final project in computational biology, we are planning on building a metagenome-assembled genome extraction tool (also called a "binning" tool).

In metagenomics, **binning** is the process of grouping DNA segments that originate in a single OTU together from an assembled metagenome. Metagenome-assembled genomes (MAGs) are very common in the field of microbial ecology, where scientists seek to characterize the genomes of organisms that cannot be individually sampled or cannot be cultured in lab. A large metagenome is sampled by collecting and sequencing an entire microbial community from an environment of interest. After the assembly of the metagenomic reads into a series of contiguous segments, the segments are clustered based on a variety of metrics. Many binning tools exist, each with a specific implementation. Generally, these tools fall in to three categories based on methodology:

1. **Cluster-based**: Algorithm uses a variety of metrics (tetranucleotide frequency, GC content, coverage) and a clustering algorithm such as affinity propagation to determine clusters of contigs, which are then taxonomically identified using a 16S rRNA annotation tool or something similar. Phylogenies from different clustering metrics are either combined after the clusters are finalized or during the assignment of clusters.

2. **Alignment-based**: The tool aligns metagenomic reads against a pre-defined protein database (such as NCBI's non-redundant protein database) and assigns each read to a taxa, and then resolves the taxa based on computed thresholds.

3. **Hydrid**: Hydrid tools use a combination of these techniques, often with alignment scores as the metrics for clustering.

2 Our plan

Because the alignment-based techniques are often database-intensive and we have limited experience manipulating databases, we will write build an algorithmic tool that uses a clustering technique. It will take as input a pre-assembled set of metagenomic contigs and metagenomic mappings (as .fasta and .bam files respectively) created with IDBA-UD (a metagenome-specific assembly tool) and bowtie2 (a mapping tool) and cluster the contigs based on the following three metrics:

1. GC content: the percentage of bases in a sequence that are either guanine or cytosine.

2. Tetranucleotide frequency: the relative frequency of each 4-base combination (AAAA, AAAT, AATT) in a genome.

3. Coverage: Presumably, in a particular environment, each microbial community has a unique population size and therefore a unique average coverage across it's contigs.

Each of these parameters are unique to a particular genome. After clustering the contigs based on each method, we can combine the sets of clusters to a set of clusters that best represents each of the cluster sets produced by our three parameters. A common clustering method is affinity propagation, which is more efficient to implement than many hierarchical clustering methods, and does not require a pre-determined number of clusters. Combining the sets of clusters for each of the three parameters may be accomplished via a naive method, such as simply adding together neighbor matrices for each cluster set, or a more complex method for combining clusters, like Evidence Accumulation. After genomes are assigned, we can attempt to taxonomically identify the genomes using the SILVA database or something similar, and create a phylogeny from this assigned taxonomy.

We will implement the methods for calculating parameters of the contigs, the clustering method, and a naive approach for creating consensus from individual cluster sets. We will write the parsers for reading input files and writing creating output files, as well as any code for assessing the quality/completeness of the final bins.

This plan has a reasonable amount of flexibility in terms of time. To assess the quality of our final binning, we can compare the sets of contigs to bins computed using hybrid tools such as anvi'o (which has a great blog). Through research with Rika, I have access to 15 large, high-quality paired-end metagenomic sequence datasets that are publicly available, as well as mapping and assembly files for the metagenomic datasets. We also may access the TARA Oceans Dataset, which has dozens of metagenomes of various sizes for testing the algorithm. If we begin to run out of time, we can cut out a complicated algorithm for combining the cluster sets, annotating/taxonomically identifying the genomes, and assigning phylogenies, simply outputting a set of fasta files for the bins. The assessment of our algorithmic tool can be as simple as answering the question "Are contigs A and B assigned to the same bin by our tool? Are they assigned together by other algorithms?" or as complex as comparing phylogenies, completeness, and taxonomic identity of the finalized bins. Sorry for the long proposal.

References

[1] Ana L N Fred and Anil K Jain. "Combining Multiple Clusterings Using Evidence Accumulation." In: (). URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.9937%7B%5C&%7Drep=rep1%7B%5C&%7Dtype=pdf.

[2] Elaina D Graham, John F Heidelberg, and Benjamin J Tully. "BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation." In: (2017). DOI: 10.7717/peerj.3035. URL: https://peerj.com/articles/3035.pdf.

[3] Marc Mézard. *Passing messages between disciplines*. Sept. 2003. DOI: 10.1126/science.1086309. URL: http://www.ncbi.nlm.nih.gov/pubmed/14500972.

[4] Tarini Shankar Ghosh, Monzoorul M Haque, and Sharmila S Mande. "DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences." In: *From Asia Pacific Bioinformatics Network (APBioNet) Ninth International Conference on Bioinformatics* (2010), pp. 26–28. DOI: 10.1186/1471-2105-11-S7-S14. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2957682/pdf/1471-2105-11-S7-S14.pdf.