Info 690: Bioinformatics for Microbial Genomics and Metagenomics

Assignment IV

Anoop Mayampurath

1) Chatterji et al. *CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads* (RECOMB 2008)

This paper introduced CompostBin, a software tool to bin sequence raw reads into taxonomic bins. Briefly, the algorithm behind CompostBin creates a feature matrix that is indicative of k-mer frequency distributions, projects the matrix on to a lower dimensional space using PCA methods, and then at attempts to classify the sequences into two bins by first building a graph and then using a semi supervised normalized cut algorithm. The process is repeated for division into multiple bins. Here are some notes

- Compost Bin provides an advantage in the fact that it does not require any training or any closely related reference genome for alignment. By performing clustering in a lower dimensional principal component space CompostBin is not subject to noise or redundancy effects that is present in the higher dimensional sequence reads.
- It is not clear how the number of principal components is finally decided on. The authors use both two and three principal components. Typically, the number of principal components is chosen based on x% percentage of the largest eigen value. The eigen value matrix indicates contribution of each principal component, with the first eigen value contributing the most, the second the next –most and so on. Eigen values within some percentage of the first eigen value are retained, and the corresponding number of principal components are chosen.
- The authors use 6-nearest method for creating a graph, 6 being chosen based on test. This is what makes the algorithm semi supervised since information from 31 phylogenetic markers is incorporated for accuracy purposes in defining the 6 nearest neighbors. Again, instead of setting this parameter to be hard-coded, a distribution of the distances between phylogenetic markers may be used for a more accurate threshold. This may prevent over fitting.
- The authors show accuracy in simulated reads acquired from utilizing ReadBin on known genomes to simulate Sanger sequence reads. The abundance values are varying. The misclassification error is appreciably low, although the results are not comparable across datasets since accuracy for some datasets is at species level while others is at a phylum level.
- The tool was also tested on publically obtained metagenomic data for gut bacteriodetes in glassy-winged sharpshooter, that previous studies had given three taxonomic bins – *B. cicadellinicola, S. muelleri* and unclassified. However, the third category was left out while testing; this is slightly suspicious and highlights one possible flaw in the tool in that since the cut always bisects the graph into two, overfitting might occur.
- Also, CompostBin requires the specification of number of bins that is expected in the sample. Thus this tool makes detection of novel species difficult. This was observed during last week's presentation on PhymmBL.

2) Zhou et al. *Barcodes for genomes and applications.* BMC Bioinformatics 2008

This paper expands on the idea of using k-mer frequency distributions that was introduced in the CompostBin, as a means of assigning signatures to genomes. The genome is divided into M base pair fragments and for each fragment, the frequency count of each k-mer occurrence is recorded. Here, M is set to 1000 and k is set to 4, although studies are done with different values of M.

The frequency count matrix is converted into a gray scale image using a mapping procedure. It is noticed that different species do indeed have different barcode images. Comparison against a randomly generated 0^{th} order Markov Chain model shows distinct differences. The images closely represents a 3^{rd} order Markov Chain model, which is understandable because of codon usage.

On comparing 4-mer frequencies between prokaryotes and eukaryotes, two overlapping but distinct distributions were observed. Comparison of barcode distance against similarity amongst 16s rna sequences shows that they are both proportional after a very high level of similarity (> 0.8). Lower similar sequences show the same level of barcode distance distribution. This could indicate that while barcode distance is a useful measure of clustering together sequences in the same species, it might not have enough discriminatory power to differentiate between closely related species and more distant species. Discrimination among family types is apparent from the scatter plot figure, so this might be a good first step in metagenomic analysis.

On developing a binning algorithm, the accuracy decreased with decrease in M and increase in number of taxa. Thus, this might not work for shorter fragments without contig assembly. There might be interesting applications to gene finding since the images showed appreciable difference between repeats, promoters, coding and non coding regions.

The images do show a distinct difference; however on reducing the difference to a distance metric as described in the paper, it is felt that resolving power might have been lost. Perhaps another metric would work better.