Heewook Lee
I609 Metagenomics
HW4
03/02/10

Summary of "CompostBin: A DNA compositino-based algorithm for binning environmental shotgun reads" by Chattergi, *et al.*

This paper introduces a novel metagenomic binning tool based on PCA method applied to k-mer frequency of samples. Given metagenomic sequence reads, CompostBin tries to distribute them into taxon-specific bins.

The motivation of the paper was similar to other metagenomic data classification/binning tools. First motivation was that binning/classification is the first step that should be conquered in order to make further inferences given metagenomic data. Their second motivation, which was targeted specifically to overcome current technique's weakness, was that they believed that successful binning algorithm should be able to perform without requiring training on currently available genomes. They claimed that the selection of currently sequenced genomes was a biased set rather than a representative set of the actual population of bacteria. Third motivation was to successfully bin short sequence fragments (short here is Sanger by the way).

They claimed that the poor performance of many other binning tools that uses k-mer frequencies was that the measuring k-mer frequency requires a feature vector of a large dimensionality, hence resulting in the associated noise. In order to cut the noise arising from the high dimensionality, they employed a popular technique, Principal Component Analysis (PCA) to project the feature vector onto a lower dimension space. Instead of using all the component, which is like a single dimension of the projection, they showed that only using the first three component suffices to achieve reasonable performance. After PCA was applied, the normalized cut clustering algorithm was employed with phylogenetic information as a guide in order to actually distribute the data into taxon-specific bins.

To demonstrate the performance, they tested CompostBin on two types of data. First, they tested on the various data sets generated by sequencer simulator, ReadSim. Second, they tested the algorithm on a publicly available metagenomic data, sequence reads obtained from gut bacteriocytes of the glassy-winged sharpshooter. The error percentage ranged from 0.21% to 8.01% in 13 data sets and was less than 6% for 11 data sets. They claimed that two data sets resulted in > 6% error was due to the small phylogenetic distance. For the publicly available metagenome data, their calculated error % was 5.9%. With reasonably low error percentage, they claimed that the algorithm is able to process the actual metagenomic data. They didn't quite explained how they calculated the error % for the real data. Since there isn't really a ground truth for the real data, I thought they should have provided such information.

This paper was written in 2007 and I am wondering why they used Sanger sequences. I thought it was odd for them to refer Sanger sequences as short sequences. This brings me to wonder how CompostBin will actually perform on real short reads generated by 454, solexa, etc.

Summary of "Barcodes for genomes and applications" by Zhou, *et al.*

This paper introduces a novel scheme of visualizing *k*-mer frequencies of sequence fragments for large scale sequence such as genome. They claim that majority of its short sequence fragments have highly similar barcodes within a genome, where barcodes are vritually drawn based on the *k*-mer frequencies. Their main motivation was from the following question: "does each genome have a unique signature imprinted on its short sequence fragments?"

For each fragments of a given size in a given sequence, they calculate the combined frequencies of *k*-mer and its reverse complement. This leads to generating a matrix of columns representing all possible *k*-mers and its reverse complements and rows representing each fragment of a given size. Given this matrix, grey-level image barcode is generated by mapping frequencies, assigning darker grey levels to lower frequencies and vice versa.

Authors claim that these grey scaled image, in overall, gives a consistent vertical bands within a genome, indicating that that *k*-mer frequency distribution is stable throughout the genome. However, they found that there are small fraction of these fragments that gives rise to horizontal bands within grey scale images, namely barcodes. These horizontal strips indicate fragments that don't have a consistent *k*-mer frequency distribution, which they define it as "*abnormal."* They have also observed that multiple chromosomes of same organisms tend to have similar barcodes and so do phylogenetically relatively close genomes. They also concluded that random sequences that are generated by a third-order Markov chain model resembles barcodes of real genome sequences.

They claimed that different classes of genomes have unique characteristics in their barcodes. I think this was somewhat expected as *k*-mer frequencies have been used widely to distinguish different classes of sequences. I thought this was somewhat absurd that they sounded as if they were claiming a novel finding. It is novel that they came up with a new scheme of visualizing the entire *k*-mer frequencies throughout genomes. However, I am not so sure if they can claim that barcodes provides a novel mean to distinguish different classes of sequences.

One application of this barcoding scheme was in identifying horizontally transferred genes. They examined those horizontal banded fragments - abnormal sequence fragments and concluded that a significant portion of them come from horizontal gene transfers, phage invasions and highly expressed genes. However, they could not identify 70% of these abnormal fragments. With 70% of these sequence fragments being unknown of origins, I am not sure whether they can claim the barcoding scheme as an effective of distinguishing horizontally transferred genes.

Another application of the barcodes was binning metagenomic sequences. They clustered the sequences and assigned into bins based on the barcode similarity function they introduced in the paper, which is a generalization of the Euclidean distance between two vectors of the averaged *k*-mer frequencies across each genome. The reported accuracy of binning compared to the previous method, PhyloPythia indicated that their methods consistently perform better. They claimed their tool as more general one since it does not require training like PhyloPythia.

I think the barcoding scheme itself is quite nice in a way that one can see the *k*-mer frequencies in one image and compare it with barcodes of other genomes; however I don't think it deserves nothing more than that because *k*-mer frequencies have been used alot to distinguish different classes of sequences.