Name: Vikas Rao Pejaver **Username**: vpejaver

Paper 9: CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. Chatterji et al. *RECOMB*. 2008

This paper discusses a new algorithm for the binning of DNA sequences obtained from environmental samples and its application to taxonomic classification in metagenomics. Like most previous methods, the algorithm is based on DNA composition features and uses k-mer frequencies as the basis for the identification of unique genomic 'signatures'. However, unlike previous methods, the highlight of this algorithm is that it can be applied to reads directly, without prior assembly or training. Thus, it avoids errors due to training data biases and possible mis-assemblies.

CompostBin is a PCA-based algorithm where the high-dimensional feature matrix (hexamer frequencies) is reduced into a low-dimensional space by principal component analysis. This transformation is particularly useful in the case of short reads where there is a lot of noise associated with high-dimensional hexamer frequencies. After PCA, the projected feature matrix is then subjected to a normalized cut clustering algorithm that first converts the matrix into a graph, based on normalized Euclidean distance. This graph is then updated iteratively based on phylogenetic marker data provided. Finally, the normalized cut algorithm cuts the data into two separate bins so that the normalized cut value is minimized. For $k > 2$ bins, this bisection is recursively repeated till k is reached.

The authors report less than 6% errors in 10 of the 12 synthetic datasets analyzed. The datasets were constructed based on random 'reads' obtained from random genomes to simulate both low- and medium-complexity samples. In order to study the accuracy of CompostBin on real data, it was tested on reads obtained from gut bacteriocytes of the glassy-winged sharpshooter. For this dataset, the error was found to be 5.9%. CompostBin was also found to be effective in binning sequences, even in cases where they came from two species from the same genus. This is impressive for a DNA composition-based method as same-genus species tend to have relatively similar DNA composition features.

An interesting point to note is that, although the authors motivate the need for CompostBin by developing their method for short reads, the tests conducted were on data simulated to mimic paired-end Sanger sequencing reads. In the context of the present day scenario, it would be interesting to see how this method performs on the short reads obtained from next-generation sequencing methods. Another key issue that needs to be clarified is that CompostBin is mainly a binning algorithm. The taxonomic classification part of the algorithm is semi-supervised. If one needs to extend its application to the classification of reads, then one has to provide label information based on phylogenetic marker genes. This seems to be a limitation for actual classification as, in many cases, this information may be unavailable. Moreover, the availability of such information depends on the goals of the metagenomics project. Another potential drawback is that CompostBin assumes that the number of bins is known and can be obtained from the label data implicitly. This may be a limitation when considering samples of much higher complexities.

As such, CompostBin offers a strong alternative method for sequence-binning. The facts that it does not depend on any pre-processing steps and that its run-time is bound by O(NK(log N)) (where K = number of bins) makes it a good tool in environmental sequencing. However, as the authors acknowledge, more improvements need to be made and some of the points mentioned above should be addressed in order to exploit its full potential. But the question is: Given the current state of the art, would we want to go back to improving CompostBin or look at some of the newer tools for newer sequencing technologies?

Paper 10: Barcodes for genomes and applications. Zhou et al. *BMC Bioinformatics*. 2008

This paper seeks to provide a unique signature to a genome so that sequence fragments can be identified from a mixture. The concept that is introduced in the paper is that of a genomic 'barcode' which provides a concise summarized image of the whole genome and also serves as a feature unique to the genome. This barcode is generated based on the frequencies of k-mers and their reverse complements over a fixed number of fragments from a genome. Genomic barcodes are consistent not just for prokaryotes but can akso be obtained for genomes across the tree of life. Furthermore, the authors also illustrate the application of the barcode concept to several problems in metagenomics and genomics. Binning of reads from metagenomics samples can be achieved by measuring similarities of barcodes of each read. Horizontally transferred genes can also be detected by identifying locally varying regions in the barcodes of genomes. The basic idea is that these variations are not inherent features of the genome and may have been acquired through horizontal transfer.

In this paper, all genomes were typically divided into fragments of length 1000 bp and tetramers were used to create barcodes. A genomic barcode can be described as a matrix where the number of rows is equal to the number of fragments and the number of columns is equal to the number of unique combined tetramers. Each entry in the matrix corresponds to the frequency of the particular tetramer in the particular sequence fragment. These frequencies are then mapped to grey levels such that the grey level has the same meaning across different genomes. By convention, darker grey levels indicate lower tetramer frequencies. For the application of these barcodes, the authors have proposed a barcode distance measure that is virtually the Euclidean distance between two vectors (barcodes). This measure is used as the basis for clustering by the CLUMP program and k-means methods (combined), during metagenome binning. In order to detect abnormal fragments in genomes, a sharp transition point is identified on the convex curve of a function that represents the fragments containing k-mers at frequencies, from the tails of the frequency distribution.

The authors deduce that the barcode method works really well for the creation of unique identifiers that provide an intuitive and informative global view for genomes. It is so often the case that vertically consistent bands are observed on the grey-scale barcodes – implying that bar-coding is a possible inherent feature. The authors attribute this to a third-order Markov chain property of coding regions in genomes. This study also justifies the use of barcodes, not merely as visualization methods, but as tools for genomic analyses as well. Barcode distance-based clustering of sequences on simulated metagenomics datasets have been shown to be quite accurate for different 'read' lengths. In fact, binning based on barcodes seems to outperform established methods such as PhyloPythia on these datasets. In the context of abnormal fragment detection, one of the highlights is that the average percentage of fragments with abnormal barcodes in prokaryotes is 12.85%. A majority of these cases cannot be explained as phage transfers, horizontally transferred genes or highly expressed genes. This opens up the potential for the discovery of novel foreign fragments in genomes.

Genomic barcodes are an interesting concept in terms of intuitive visualization and analysis. But the question is: would it make sense to use these 'barcodes' as the basis for metagenome binning? After all they are just feature vectors of k-mer frequencies and add very little to existing methods in binning. In that aspect, barcodes seem like a fancier way of doing the same things that DNA-composition based binning algorithms do. Moreover, by not performing tests on real datasets, it remains to be seen if barcodes can really be effective for the binning of metagenomic sequences, especially when they come from phylogenetically closed species. As such, the concept of genomic barcodes, in itself, is very interesting but it does not seem like the most intuitive solution to the binning problem.