Chad Burrus

I609, Summary #4

February 27, 2010

Zhou, Olman, and Xu. *Barcodes for genomes and applications.* BMC Bioinformatics, 2008, 9:546.

Zhou, Olman, and Xu worked on a process they named "barcoding" for genomes that sounds like GC content profiling on steriods to me. Essentially, they broke the genomes they worked with into blocks of fixed sizes and counted the occurrence of each *k*-mer and its reverse complement in each individual genome. They then sorted the *k*-mers by frequency and split the *k*-mers into subgroups. Using that ratio, they assigned gray levels to each *k*-mer, then assembled barcodes for the blocks by making gray dots for the *k*-mers in alphabetical order and aligning the blocks from first to last. Then when you look all of the blocks as a whole, the blocks form a barcode-like pattern (with some very special exceptions). That overall pattern appears to be (a) unique to each genome, (b) varied enough to separate coding regions and non-coding regions, (c) similar to related species relative to how related the species are, and (d) distinct enough to tell apart different types of genomes, allowing four types of information to be obtained from these barcodes.

The exceptions to the overall barcode pattern are pretty special, indeed. They tend to come from three classes: horizontally transferred genes, genes from bacteriophage invasions, and highly expressed genes. Considering the overall uniformity of the barcodes in the rest of the genome, it's relatively easy to understand why the horizontally transferred and bactiophage genes are quite different--they come from a different organism originally and haven't been around long enough to fit in with the rest of the genome that well. The highly expressed genes are less obvious, and one potential answer I thought of was that they are highly conserved as well, and so potentially have a very ancient origin before the genome specific pattern formed.

These similarities allowed the authors to develop a new binning algorithm to sort metagenomic reads into their various species. The algorithm works by computing the barcodes for the genomes, then calculating the "distance" between them. It works far better than PhyloPythia (though that's not saying much), able to classify reads into the correct species >50% of the time even at read lengths of 500 bases while PhyloPythia can only get the genus right 45% of the time at read lengths of 1000 bases, though admittedly on different data sets. Overall, though, its numbers are close to or a little better than the theoretical limit on binning accuracy.

Chatterji, *et al*. *CompostBin: A DNA composition-based algorithm for binning envirnonmental shotgun reads. Research in Computational Molecular Biology*, Vol. 4955 (2008), pp. 17-28.

The authors worked a new method of binning reads from environmental shotgun sequencing. Their method, called CompostBin, works directly on the raw reads instead of the contigs that other methods use so the assembly process can't skew the reads. It was prompted by the authors' belief that just viewing metagenomic data as a "bag of genes" without regard to species composition was not a good solution and seeks to address the two major limitations of previous methods.

CompostBin is a composition-based algorithm that does not require a training set to classify reads with a high degree of accuracy. Instead of trying to look at everything at once, CompostBin uses a PCA technique to reduce the number of variables we need to look at (currently to 3), then applies a normalized cut clustering algorithm with phylogenetic marker processing to do the final classification. If more than two bins are required, it moves through the list of current bins and repeats the PCA/clustering process on the one most likely to divide easily until we have the number of bins we need If none will divide easily, the process ends even if the desired number of bins has not been reached.

CompostBin tends to work well, as mentioned before. It was tested against 12 simulated datasets of various composition and complexity, and against one real sample where the composition has been determined. In 11 of their 13 datasets, the number of mis-classified reads is less than 6%, and in the two cases where it's not, the species being compared are closely related. However, closely related species can be distinguished if they are divergent enough, as evidenced by the test with the real metagenome.

The authors also have several ideas in mind to help improve the results further, including changing the way they reduce variables, dynamically choosing the most appropriate *k*-mer to use, and using more efficient data structures and algorithms. One possible problem with CompostBin is that it requires the number of bins to be set up front--while this isn't hard to compute (16S rRNA tests or similar will do nicely), it's an extra step that needs to be done that I think the authors were trying to avoid--at least it sounds like they were trying to build a tool that you can just point at your data and let it work.