THE INTRINSIC BARCODE FEATURE OF A GENOME AND ITS APPLICATIONS

Fengfeng Zhou, Victor Olman, Xizeng Mao and Ying Xu *
[1] *Computational Systems Biology Laboratory,*
Department of Biochemistry and Molecular Biology, and Institute of Bioinformatics,
University of Georgia, Athens, GA 30602, USA,
[2] *BioEnergy Science Center, http://genomicsgtl.energy.gov/centers/center_ORNL.shtml*
**Corresponding author: xyn@bmb.uga.edu*

Background: A K-mer is a short nucleotide sequence of length K. It is known that the frequency of any K-mer within the fixed-length windows along a genome is largely a constant. But little is known about why a genome is organized to have this property.

Objectives: We believe that this is an intrinsic feature of any genome and is worth of further studies. Specifically, this feature will help us understand the evolutionary mechanism of the nonprotein-coding regions in a genome, which remains largely unknown.

Methods: We recently developed a mapping strategy from the frequencies of all K-mers to a grayscale picture within a fixed-length window along a genome (F Zhou, et al., BMC Bioinformatics 2008 9:546). The gray scale, mapped from K-mers frequencies, remain approximately a constant across all the windows along a genome. So the gray-scale image of a genome consists of many lines, resembling a commodity barcode. So we call such an image of a genome its barcode, and the database BoDB (http://csbl.bmb.uga.edu/~ffzhou/BoDB/) lists the barcodes of all the sequenced archaeal, bacterial, eukaryotal, plastid and mitochondrial genomes. We have studied why a genome is organized in such a way to have the barcode properties, and have carried a number of applications of these properties on classifications of functional elements in a genome.

Results: It is interesting to find that a genome's barcode is essentially the same to the barcode of a random sequence generated using the 3rd order Markov chain models, while in the barcode of each genome, there could be fractions of the genome with substantially different barcodes with those of the rest of the genome, which we have found in general correspond to horizontally transferred genes. By utilizing this feature, we have applied the barcode property to identify horizontally transferred genes in a genome, as well as to bin the metagenomic fragments so that fragments in a bin tend to be from the same taxonomic group. Other applications of the barcode properties include prediction of target genes of microRNAs and classification of transposons.

Significance of the study: The barcode of a genome should be the result of the interplay of the complex life system. The data suggest that nucleotide elements in the same genome tend to have similar barcodes, and a newly inserted nucleotide element will become more and more similar to the host genome over the time.