DNA Symphony: A new method to represent genomic sequences

Rosario A. Medina Rodríguez[1], Harieth M. Bernedo Cordova[2], Jesús P. Mena-Chalco[3]

[1] University of São Paulo, São Paulo, Brazil, [2] San Pablo Catholic University, Arequipa, Peru, [3] Federal University of ABC, São Paulo, Brazil

rmedinar@vision.ime.usp.br, hariethbc@gmail.com, jesus.mena@ufabc.edu.br

Contribution

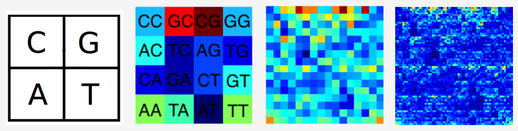
We propose a new method for representing DNA sequences by mapping the k-mers frequencies extracted from the genomic signature of different genomes into a synchronized polyphonic musical composition.

Unlike the existing methods of DNA audio representation, our method:

- Represents the main patterns and organization from the complete genome as it uses its genomic signature to be translated into musical notes;
- Considerably reduce the length of the music clip by using a vector of frequencies depending on the k-mer size instead of the genome length;
- Creates a polyphonic track from different genome sequences which is analogous with the alignment of them.

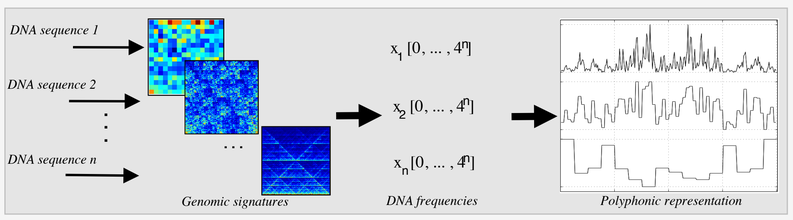
Genomic Signature - FCGR

Many graphical methods for DNA representation have been reported in the literature. These methods provide a simple way of viewing, storing and comparing many sequences. The Chaos Game Representation of Frequencies (FCGR), estimates information for each possible DNA word with fixed size (k). The result is a matrix ($2^k \times 2^k$), called "genomic signature".

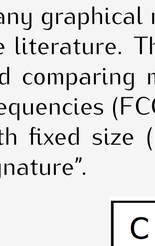


(a) *k-mer: 1* (b) *k-mer: 2* (c) *k-mer: 4* (d) *k-mer: 6*

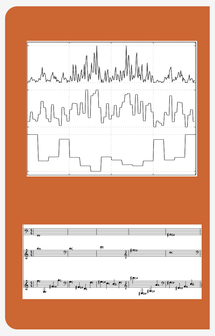
Method: DNA Symphony



Generate a genomic signature for each genomic sequence (both strands). Normalize those values between 0 and 255.



Translate the genomic signature into a vector of 4^n length by reading the matrix in a "U-inverted" way. Finally, assign sound frequencies and based on experimental results we chose the following parameters:
(*i*) range of 35 to 85 Decibels (Db), to avoid too low/high notes;
(*ii*) note duration of a Crotchet and
(*iii*) velocity equals to 64.



Generating polyphonic representation of $\mathcal{S}$ genomes. We create a polyphonic audio with $|\mathcal{S}|$ genomes, each one in a different channel.
When the k-mer sizes are different, the notes should be synchronized according to the largest k-mer size.
In this context, notes generated from a genomic signature with small k-mer size will have a proportional duration to the ones with bigger size.

Experimental Results

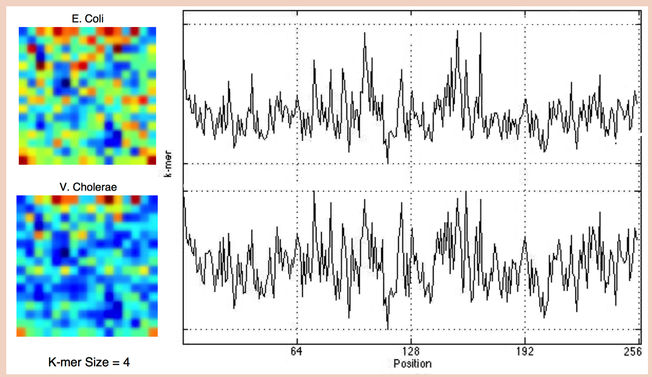
Different k-mer sizes

- Instrument: Piano; Number of Channels: 6;
- Genomic Signature (k-mer sizes): 1, 2, 3, 4, 5, 6 and 7;
- Species: *Solanum tuberosum*, *Escherichia coli*, *Puma concolor*, *Mycobacterium leprae*, *Secale cereale* and *Clostridium acetobutylicum*.

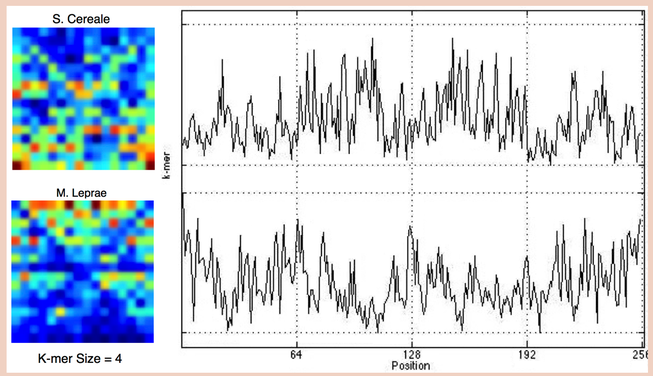
Same/Different families

- Instrument: Piano; Number of Channels : 2;
- Genomic Signature (k-mer size): 4;
- Species: *Escherichia coli*, *Vibrio cholerae* and *Secale cereale*, *Mycobacterium leprae*.

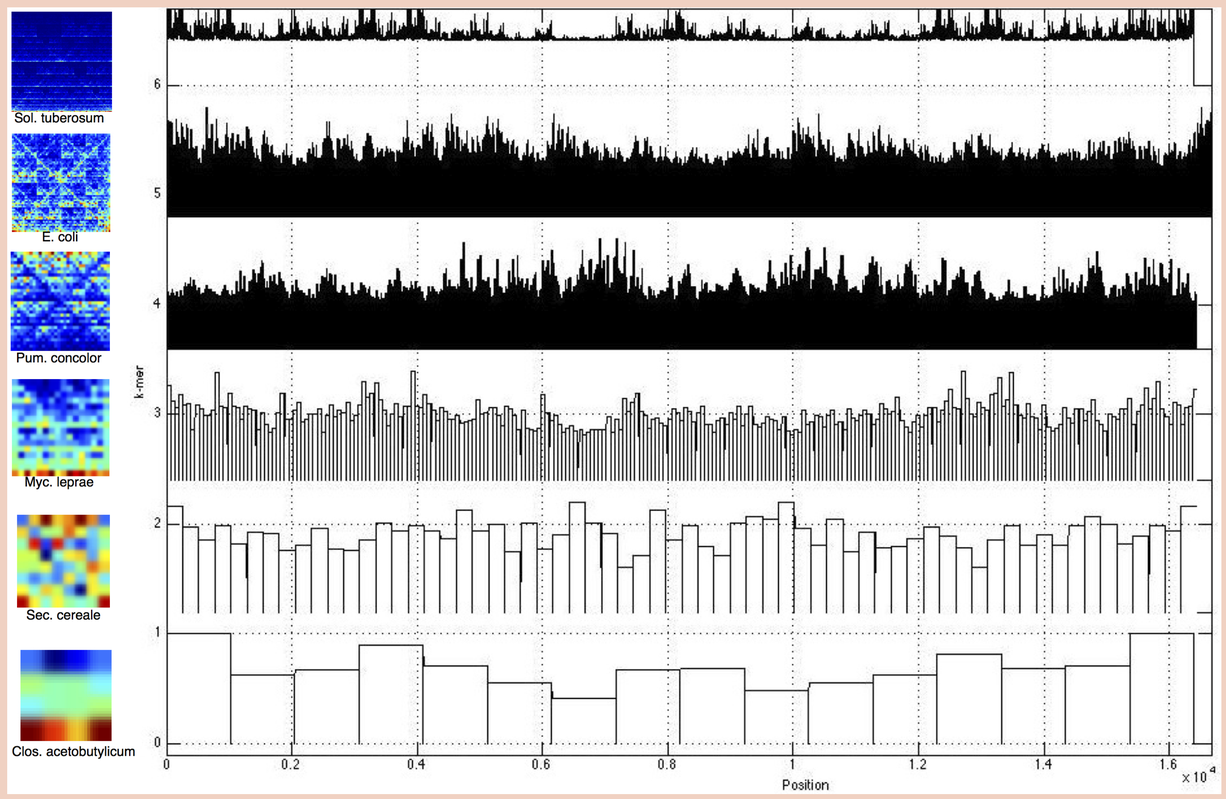
Same Family - *Bacteria*



Different Family - *Eukaryote/Bacteria*



Different k-mer sizes



Conclusions

This method generate a polyphonic audio sequence composed by a set of DNA sequences, which preserves the structure and organization of the original genomes. Based on the experiments, as it was expected, the mapped values from similar genomic signatures have similar audio sequences. Thus, when they played in different channels present similar patterns along the audio sequence.

This new representation could be used to create genomic signatures that represent a whole family of species. Future works are devoted to validate our method by applying multiple sequence alignment.

More information and audios: http://www.vision.ime.usp.br/~rmedinar/DNASymphony/