DNA Symphony: A new method to represent genomic sequences

Rosario A. Medina Rodríguez
University of São Paulo
São Paulo, Brazil

Harieth M. Bernedo Cordova
San Pablo Catholic University
Arequipa, Peru

Jesús P. Mena-Chalco
Federal University of ABC
São Paulo, Brazil

***Abstract*—Complete genomic sequences from biological species can be graphically represented by a “genomic signature”. This representation provides information about the oligonucleotide frequencies considering different size of k-mers. Moreover, genomic sequences can also be represented by an audio signal, obtained by translating each oligonucleotide or protein into a certain range of audio frequencies. Although audio representation strategies provide an interesting result, they only use part of the genomic sequence. To date no method exists which contemplates the complete genome sequence. This work proposes a new method for audio representation of genomes by composing a polyphonic signal using a set of complete genomic sequences. This method is described here by first extracting the genomic signature for each sequence. Then, to obtain the audio signal, two-dimensional genomic signatures are transformed into a one-dimensional sequence by normalizing each value into an audible spectrum. Finally each signal, depending on the number of sequences, is played on a different channel to generate a polyphonic track. The experimental results and the audio analysis suggest that the described method, preserves the main patterns and genome structure from the original sequence.**

I. Introduction

Biological data has experienced a significant growth in the number of available genomic sequences from different organisms [1]. In addition, with the advances of new technologies for sequencing, the need for efficient algorithms and analytical tools for genome representation and processing is increasing continuously [2]. In the past years, 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.One of the most commonly used method is Chaos Game Representation of Frequencies (FCGR) [3], which estimates estimates information for each possible DNA word with fixed size (k), known as *k-mer*.This matrix is called “genomic signature” and represents a measure of the resolution or granularity in the analysis of primary DNA sequence organization [4].

DNA sequences are not only graphically represented, it is also possible to capture some music patterns from them. One of the first attempts was proposed by Susumu and Midori Ohno [5]. They transformed DNA sequences into musical scores based on an eight note scale. In addition, weights were assigned for each nucleotide based on its molecular weight. The most recently work is based on some principles from jazz bebop improvisation. It is called Microbial Bebop [6] and uses musical concepts to highlight the relationships between multiple data types in complex biological datasets.

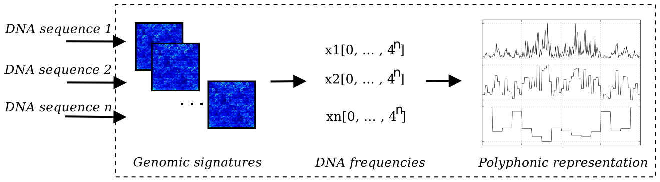


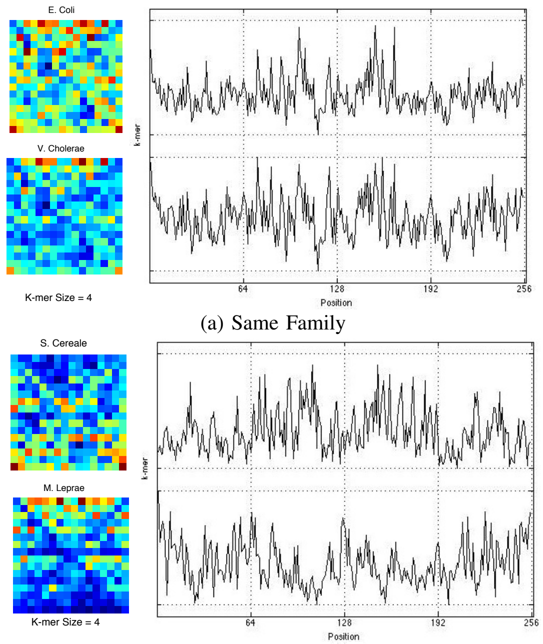
Fig. 1. Schematic flow diagram of proposed method.

Considering that the numerical characterization of the DNA sequences could be made for each nucleotide, amino acid or protein. In this paper, 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: (i) Represents the main patterns and organization from the complete genome as it uses its genomic signature to be translated into musical notes; (ii) 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; and (iii) Creates a polyphonic track from different genome sequences which is analogous with the alignment of them.

II. Method: DNA Symphony

Given a set of genome sequences $\mathcal{S}$, the methodology we followed is represented in Fig. 1 and described in the items bellow:

1) *Generating genomic signatures:* Firstly, we generate a genomic signature for each genomic sequence (matrix of frequencies). Then those values are normalized between 0 and 255. It is worth noting that both strands of the DNA sequence were used in this work.
2) *Mapping frequencies into an audio sequence:* Secondly, we translate the matrix of frequencies into a vector of 4^n length. The matrix is read from the bottom left corner, through the top left corner until the bottom right corner (“U-inverted” way). This way of reading allows the recovery the original DNA sequence from a genomic signature. Finally, to assign sound frequencies and based on experimental results we chose the following parameters: (i) range of 35 to 85 Decibels (Db), in order to avoid too low/high notes; (ii) note duration of a *Crotchet* and (iii) velocity equals to 64.
3) *Generating polyphonic representation* As the final step, we propose the composition of a symphony



(a) Same Family

(b) Different Families

Fig. 2. Genomic Signatures and Signal Frequencies. Colors in genomic signatures are mapped with the frequency values, i.e. low to blue and high to red.

with the genomes in $\mathcal{S}$. Considering the maximum of sixteen channels, 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.

III. Experimental Results and Discussion

In order to evaluate the proposed methodology, we performed 3 experiments with: (i) genomes from different/same species and (ii) genomic signatures generated with different/same k-mer sizes. Each one performed to show some relationship between "species - genomic signatures - musical compositions". Data files, including all the experiment results are available at www.vision.ime.usp.br/~rmedinar/DNASymphony.

A. Same Family of Species

We use two DNA sequences: *Escherichia Coli* and *Vibrio Cholerae*, both belonging to the *Bacteria* Kingdom. The genomic signatures were generated for k-mer sizes equal to 4, played with a Piano instrument using 2 channels. The main goal is to prove that similar genetic signatures produced similar music notes which lead to a constant pattern in the polyphonic musical composition. The results are represented in Fig. 2. The first two lines represent species from the same family; both genomic signatures are similar. This can be seen in the signal frequencies on the right side, where it is easy to distinguish the same patterns on both signals. In contrast with the last two lines where the genomic signatures and signal frequencies are different.

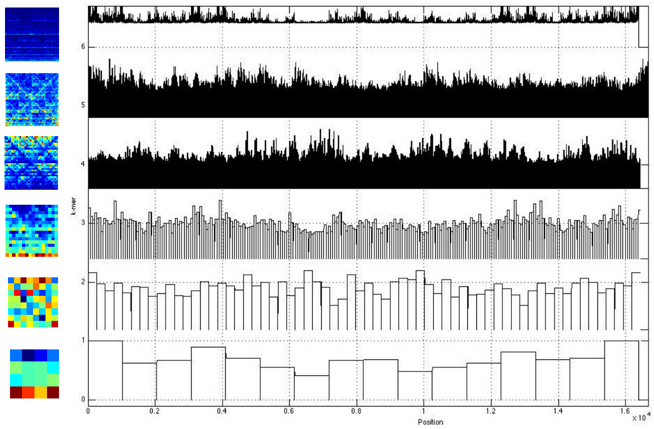


Fig. 3. Genetic Signatures and Synchronized Polyphonic Audio. Each line represent a different k-mer size.

B. Different k-mer sizes

We applied the synchronization process on 6 vectors with sizes proportional to genomic signatures with k-mer sizes $1, 2, 3, 4, 5, 6$ and 7, as can be seen in Fig. 3. The Piano instrument were used for: *Solanum Tuberosum*, *Escherichia Coli*, *Puma Concolor*, *Mycobacterium Leprae*, *Secale Cereale* and *Clostridium Acetobutylicum* species.

IV. Conclusions

This work presents a new method to represent genome sequences, based on the values extracted from its genomic signatures. These genomic signatures represent in a very effective way the main features of the whole genome. The main goal is to 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.

References

[1] A. Hey, S. Tansley, and K. Tolle, *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research, 2009.

[2] A. Roy, C. Raychaudhury, and A. Nandy, "Novel techniques of graphical representation and analysis of dna sequences—a review," *Journal of Biosciences*, vol. 23, no. 1, pp. 55–71, 1998.

[3] J. S. Almeida, J. A. Carrico, A. Maretzek, P. A. Noble, and M. Fletcher, "Analysis of genomic sequences by chaos game representation," *Bioinformatics*, vol. 17, no. 5, pp. 429–437, 2001.

[4] Y. Wang, K. Hill, S. Singh, and L. Kari, "The spectrum of genomic signatures: from dinucleotides to chaos game representation," *Gene*, vol. 346, pp. 173–185, 2005.

[5] S. Ohno and M. Ohno, "The all pervasive principle of repetitious recurrence governs not only coding sequence construction but also human endeavor in musical composition," *Immunogenetics*, vol. 24, no. 2, pp. 71–78, 1986.

[6] P. Larsen and J. Gilbert, "Microbial bebop: Creating music from complex dynamics in microbial ecology," *PLoS ONE*, vol. 8, no. 3, p. e58119, 2013.