Learning and Nonlinear Models - Revista da Sociedade Brasileira de Redes Neurais (SBRN), Vol. 6, No. 2, pp. 111-120, 2008

The self-organized chaos game representation for genomic signatures analysis

Antonio Neme, Antonio Nido
Non-linear Dynamics and Complex Systems Group, Universidad Autónoma de la Ciudad de México
{neme,nido}@nolineal.org.mx

Víctor Mireles, Pedro Miramontes
Faculty of Sciences, Universidad Nacional Autónoma de México
syats.vm@gmail.com,pmv@fciencias.unam.mx

Abstract

Genomic signatures are important as a source of comparison and classification of genomes. In particular, the Chaos Game Representation, an iterative mapping method, generates a frequency distribution of nucleotides of length *k* and represents it on a lattice of size 2^kx2^k. However, it lacks continuity in the sense that very different sequences are represented on contiguous cells of the lattice. Here, we propose an alternative method that organizes cells in such a way that continuity is higher than in the Chaos Game Rrepresentations. The cell organization is the outcome of a Self-Organizing Map and the obtained frequency distribution is named Self-Organized Chaos Game Representation. Experiments show that this visualization method is, at least, as good as the Chaos Game Representation, but it gives it a more intuitive sense when interpreting the images.

Keywords

Self Organizing Maps. Genomic Signatures. Chaos Game.

1. Introduction

Genomes are too long and too intricate to show its main features based solely in a detailed inspection of all of its components. Nucleotides are the structural molecules that constitute nucleic acids, such as DNA. The genome of an organism is the complete sequence of nucleotides, that is, the complete DNA sequence (or RNA, as in some viruses). Genomes vary in sizes from as short as a few thousands of nucleotides (or bases) as in the case of some viruses to up to several billions as in the case of some fishes [1]. There are four nucleotides in DNA: adenine (A), guanine (G), cytosine (C) and thymine (T). From an informational point of view, genomes can be considered as a sequence of these four symbols.

Because of the size of genomes, it is necessary to apply different tools to enhace some hidden features. One alternative to do so is through Genome Signatures (GS), whose goal is to become a unique and reduced (and thus, manegeable) form of the original genome [10]. The Genomic Signature is set of variables measurable over the DNA sequence that putatively tell appart different species.

Several GS have been proposed since Karlin and Burge introduced the concept [6] based on dinucleotide relative abundance. Some of the GS are based on information theory concepts [7], others rely on statistical properties [9, 8, 11, 12], and some others state that what is important is the spatial information between sequences [4, 13]. All GS proposals depend on the idea that phylogenetically close genomes have similar GS, while phylogenetically distant organisms present different GS. Also, subsequences of each genome are represented by similar GS, whereas subsequences of different genomes are different. Thus, the GS may be seen as a tool to differentiate a wide variety of organisms. A GS is an abstraction of the genome in which the important features (mainly unknown) are present.

Each genome may be seen as a sequence *S* of four letters: A, C, G, T. So, the idea of a GS is to obtain a compressed version of *S* that could be compared with other GS from different genomes. In mathematical terms, the idea is to obtain a map from a multidimensional representation to a low-dimensional representation that is more manageable [5]. A genome *S* may be

described as a single point in a multidimensional space, and its GS is achieved through a bijective function whose domain is the genome and the codomain is represented by the GS.

The Chaos Game Representation (CGR) has been identified as a GS. CGR has been proposed to visualize DNA primary structure and the resulting image to be a GS [2]. A CGR is plotted in a square with vertices labelled by the nucleotides A, C, G and T. The visualizing procedure is: Each nucleotide is plotted in the middle point from the current position *(x, y)* to the vertex representing the nucleotide (x_i, y_i) $(I \in A,C,G,T)$. The middle point is now the current position and the next nucleotide dictates the next vertex (see fig 1).

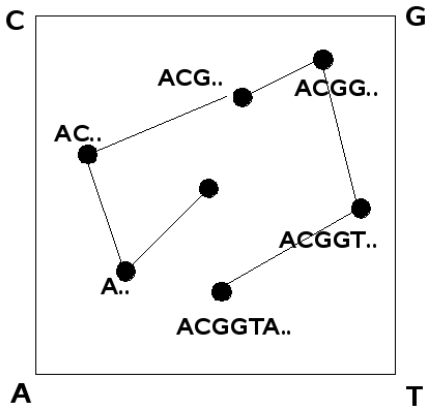

Figure 1. The Chaos Game Representation (CGR) scheme. The next nucleotide determines the vertex to whom the actual position will approach.

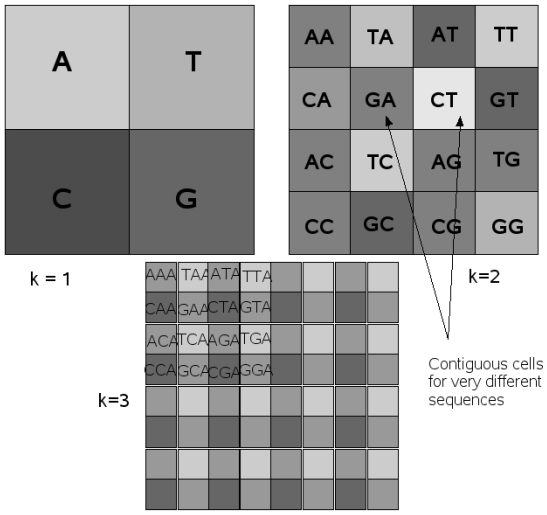

Figure 2. The Chaos Game Representation (CGR) space for length sequences of *k = 1*, *k = 2* and *k = 3*. The structure of the space is self-referenced as each new square is constructed with the original pattern. Each cell represents the relative frequency of a unique k-sequence of the genome. For *k = 2* note that there are eight pairs of adjacent cells that differ in two symbols: (TA,AT), (GA,CT),(TC,AG), (GC,CG), (CA,AC), (GA,TC), (CT,AG) and (TG,GG)

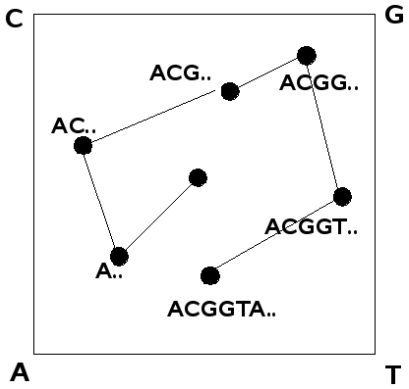

Figure 1. The Chaos Game Representation (CGR) scheme. The next nucleotide determines the vertex to whom the actual position will approach

The CGR space is recursive, as each division of the square is constructed in the same way as the original (see fig 2). This recursive process may lead to a fractal structure. The cell distribution is quite intuitive, once the recursive rule of cell distribution is understood. It is a very appropriate distribution because of the recursive structure. In this space, it is straightforward to locate the cell in which a given sequence is located [2, 5].

The lattice in which the genome is to be represented depends on the resolution k, that is, the length of the sequences from the genome that is being analyzed. The lattice is the CGR space and is defined over a $2^k \times 2^k$ cells, and each cell represents a unique k-length subsequence and all possible subsequences are represented by a single cell (fig. 2). The CGR represents the relative frequency of each possible subsequence and in general, a gray scale is applied to show this relative frequency. Highly frequent k-length sequences are represented by light gray on the corresponding cell while very infrequent subsequences are represented by dark tones. The resolution of the space is determined by k. An analysis with $k = 7$ leads to a CGR space with more resolution than that reached with, say, $k = 3$.

Although several criticisms have been directed to the CGR as a GS [3], it has proved that its features are good enough as to be a valid GS, as: **1.** the distribution of positions in the CGR space has been shown to be a generalization of Markov chain probabilities [4]; and **2.** the CGR is completely determined by all frequencies of length k sequences [14].

However, the CGR space is not necessarily the best space if a good map from the sequence space is to be considered. The reason is that the cell distribution over the grid is not ordered in the sense that the difference in contiguous cells should be minimum. That is, very different sequences may be represented in contiguous cells, as, for example, sequences GA and CT in fig 2. Also, sequences similar to each other may be represented by distant cells, as it is shown in fif. 2 with sequences AC and AG.

A map in which the relative frequencies are to be shown should have the topographic preservation property. In the sequences space, each possible sequence of length k is represented by a single point. Similar sequences are represented by close points while very different sequences are represented by distant points. Take, for example, the sequences S1 = ATTC, S2 = TAAG and S3 = ATTG. It is necessary to define a similarity measure over the four possible simbols. There are 24 possible relations, from which $A \leq C \leq G \leq T$ is one of them. What this means is that C is more similar to A than it is to T, but it is equally similar to G and A.

The idea is to obtain a lattice in which cells that represent similar subsequences of length k in the subsequence space are close to each other. The best cell distribution is that in which each cell's contiguous cells represents sequences with only one difference. Of course, when $k > 2$ this is not possible, as, if von Neumann or Moore neighborhoods are defined, the maximum number of neighbor cells is four and eight in each case while the number of sequences differing in just one nucleotide from a given k-length sequence is $k \times 3$.

We propose to use SOM to achieve a lattice with an average discontinuity between adjacent cells lower than the one in the CGR lattice. In section 1, we describe the main features of the so called Self-Organizing Map. In section 2, the Self-Organized Chaos Game Representation is introduced while in section 3 several results are presented. Finally, in section 4, conclusions as well as discussions are stated.

2. The Self-Organizing Map

The self-organizing map (SOM) is presented as a model of self-organization of neural connections, which is translated in the ability of the algorithm to produce organization from disorder [16]. One of the main properties of the SOM is the ability to preserve in the output map those topographical relations present in the input data [15],a very desirable property for data

visualization and clustering. This property is achieved through a transformation of an incoming signal pattern of arbitrary dimension into a low-dimensional discrete map (usually one or two-dimensional) and by adaptively transform data in a topologically ordered fashion [15, 17]. Each input data is mapped to a single neuron in the lattice to the one with the closest weight vector to the input vector, or best matching unit (BMU).

The SOM preserves relationships during training through the learning equation, which establishes the effect each BMU has in any other neuron. Weight neurons are updated accordingly to:

$$w_n(t+1) = w_n(t) + \alpha_n(t)\, h_n(g,t)\, (x_i - w_n(t)) \quad (1)$$

where $\alpha(t)$ is the learning rate at time t and $h_n(g, t)$ is the neighborhood function from BMU neuron g to neuron n at time t. In general, the neighborhood function decreases monotonically as a function of distance from neuron g to neuron n.

The SOM structure consists, generally, of a two-dimensional lattice of homogeneous units. Each unit maintains a dynamic weight vector which is the basic structure for the algorithm to lead to map formation. The input space dimension is considered in the SOM by allowing each weight vector to have as many components as dimensions in the input space. Each input vector x is mapped to the unit i whose weight vector is closest to it.

Since the SOM lacks resolution, that is, several input vectors tend to be mapped to the same BMU [15], a variant of the SOM was applied. In this variant, each BMU enters a refractory period in which it can not be the BMU for any input vector until the refractory state is over [18, 20]. With this scheme, the number of unoccupied cells is zero, as is in the case of CGR space. The idea behind refractory period for BMUs is that no unit maps for several input vectors so resolution may be improved.

3. The Self-Organizing Chaos Game Representation

Although SOM has been widely applied in the GS context [11, 10], it mainly has been seen as a tool for visualization of multidimensional GS. Here, we have applied SOM not only as a tool to analyze multidimensional GS, but also to construct the CGR space. As SOM is a tool that allows clustering and visualization of multidimensional data on a low-dimensional map, we applied it to construct a CGR space in which cells are ordered and thus the image representing the GS is more descriptive. Here, data is k-dimensional and there are 4^k input vectors, each corresponding to one possible sequence of length k.

The main idea is that each possible sequence of length k is represented as a vector in a k-dimensional space and a low dimensional map (two-dimensional) that represents the distribution in k-dimensional space is needed for a good visualization. Fig 3 shows a different ordering of cells for $k = 2$ obtained by SOM. The number of pair cells that differ in two symbols in this new arrangement is three, while in the CGR case (see fig 2) is eight.

CA	TA	TT	AT
AA	GA	CT	GT
AC	GC	CG	GG
CC	TC	AG	TG

Figure 3. Lattice of $2^k \times 2^k$ (k = 2) cells obtained by the SOM in which the average number of discontinuities between adjacent cells is lower than in the CGR space. Only three adjacent cell pairs show two different symbols (GA,CT), (GC,CG) and (TC,AG).

As there are four nucleotides, there are 4^k possible sequences and each sequence should be represented by a vector. As stated in the introduction, it is necessary to establish the similitude between each pair of symbols to define the sequence space. The sequence space is an abstraction of the chemical differences and similitudes between nucleotides. Nucleotides adenine and guanine are big molecules and correspond to the purines (P), whereas cytosine and thymine are part of the pyrimidine class (R) and are small molecules [1].

From these chemical properties, A and G are more similar between them that between them and C and T. In other words, $d(P,P) < d(P,R)$, where $d(a,b)$ is the distance between symbols a and b. Also, $d(R,R) < d(P,R)$.

In this work, and following from the previous discussion, we have adopted the convention that $A < G < C < T$, and coded as *0.0, 0.25*, *0.5* and *1.0*, respectively. For example, the sequence of length $k=5$ ATTGA will be coded as [0.0,1.0,1.0,0.25,0]. These coding scheme was chosen as guanine and adenine are purines and thus are similar, whereas cytosine and thymine belong to the family of pyrimidines. By these ordering, it is explicit the fact purines are different from pyrimidines, but because of chemical properties, guanine and cytosine are less different that adenine and thymine [1]. Other ordering schemes have been proposed, as, for example, those in [2, 4, 5], in which different values are assigned to each nucleotide.

The SOM algorithm was applied to the sequence space in order to obtain a topographic map in which each unit of the two-dimensional lattice becomes active (maps) for a single sequence. The size of the lattice is $2^k \times 2^k$ units, where k is the sequence length. The number of epochs for training was 1000 and the initial learning factor was settled to 0.1 and reduced exponentially to 0.0001, whereas the initial neighborhood span was 2^k and exponentially reduced to 1. The sequence space, defined by all possible sequence of length k over these four symbols, does not impacts the overall map.

Once a SOCGR space is obtained through the SOM algorithm, each genome is analyzed. For each sequence of k nucleotides of the genome, its corresponding cell is identified and the relative frequency of that cell is increased, as in the CGR model (see section 1). The SOCGR space is mathematically equivalent to the CGR space, but the cell arrangement is a better approximation of the sequence distribution of the sequence space in the sense that adjacent cells tend to map similar sequences. That is, the GS is constructed over a topographic map, while the CGR space lacks of this property.

There are many tools that may be applied to find a lattice with the minimal discontinuity between adjacent cells, as, for example, genetic algorithms. However, we used the SOM algorithm since it is not clear that an optimal algorithm may exist. The input space of all possible sequences of length k is mapped to a two-dimensional lattice in such a way that similar sequences are mapped to nearby units. At the end, SOM leads to a permutation of units in which differences among neighbors are minimized.

4. Results

In order to show that SOCGR is an efficient GS, 34 organisms were studied for subsequences of length 5, 6 and 7 (see table 1). The data was obtained from the GenBank database release 171 [19] For the human and the arabidposis thaliana genomes, two and three chromosomes were studied. Figs 4-5 show some of the SOCGRS obtained. It is also shown the CGR for the studied organism. It is important to note that the fractal structure of CGR is achieved mainly by the fact the lattice is self-similar (see sec. 1). Even though it is a highly attractive image, it is not the most continuous one. It is equivalent to seeing a set of functions in the plane whose domain has been split. The idea of SOCGR is that of a better visualization of frequencies based on an ordered CGR space. Both cell distributions, CGR and SOCGR are permutations of the same set of 4^k cells, but the later is a space with less discontinuities than the former.

No		No.	
		19	*H sapiens chr 21*
1	*A. aeolicus*	20	*L lactis*
2	*A salmonicida*	21	*M thermophila*
3	*A fulgidus*	22	*M jannaschii*
4	*A tumefaciens*	23	*M pulmonis*
5	*A thaliana chr 2*	24	*M tuberculosis*
6	*A thaliana chr 3*	25	*M genitalum*
7	*A thaliana chr 4*	26	*O sativa chr 1*
8	*B subtilis*	27	*P profundum*
9	*C elegans*	28	*P aeruginosa*
10	*C trachomatis*	29	*P troglodytes*
11	*D rerio chr 17*	30	*P aerophilum*
12	*D radiodurans*	31	*S enterica*
13	*D melanogaster chr 2*	32	*S solfataricus*
14	*E coli 32*	33	*S aureus*
15	*G gallus chr 12*	34	*S tokodaii*
16	*H salinaris*	35	*T maritima*
17	*H pylori*	36	*T volcanium*
18	*H sapiens chr 3*	37	*V cholerae*

Table 1. Analyzed organisms. There are bacteria, archaea and eukaryotic chromosomes.

To obtain the number of discontinuities, D, between a cell i and its neighbors cells h_i, it is necessary to identify the sequence mapped into i and the sequences mapped to h_i, named s_i and s_h, respectively. The sequences are of length k and the mismatches between each position in s_i and s_h are taken into account:

$$D(i,h_i) = \sum_{j=1}^{k} m(s_i(j), s_h(j)) \quad (2)$$

where $m(a,b) = 0$ if $a = b$ and 1 otherwise.

The average number of discontinuities in both, the CGR space and the SOCGR space are shown in table 2 for $k = 5$, $k = 6$ and $k = 7$. It is observed that the number of discontinuities, defined as the average number of different symbols between sequences mapped to neighbor cells (in the sense on von Neumann), is lower in the SOCGR.

Discontinuities are not the only error measure for the sequence maps. Distances between mapped sequences are also important. As in the case of discontinuities, sequences mapped to each cell have to be identified but now what is considered is not if a mismatch is present, but the distance S between the corresponding vectors:

$$S(i,h_i) = \sum_{j \in s_h} d(s_i - s_h(j)) \quad (3)$$

where $d(a,b)$ is the distance between vectors a and b. In table 2, it is shown that the distance between sequence in adjacent cells is lower in the map obtained by the SOM than the distances for the case of the CGR space.

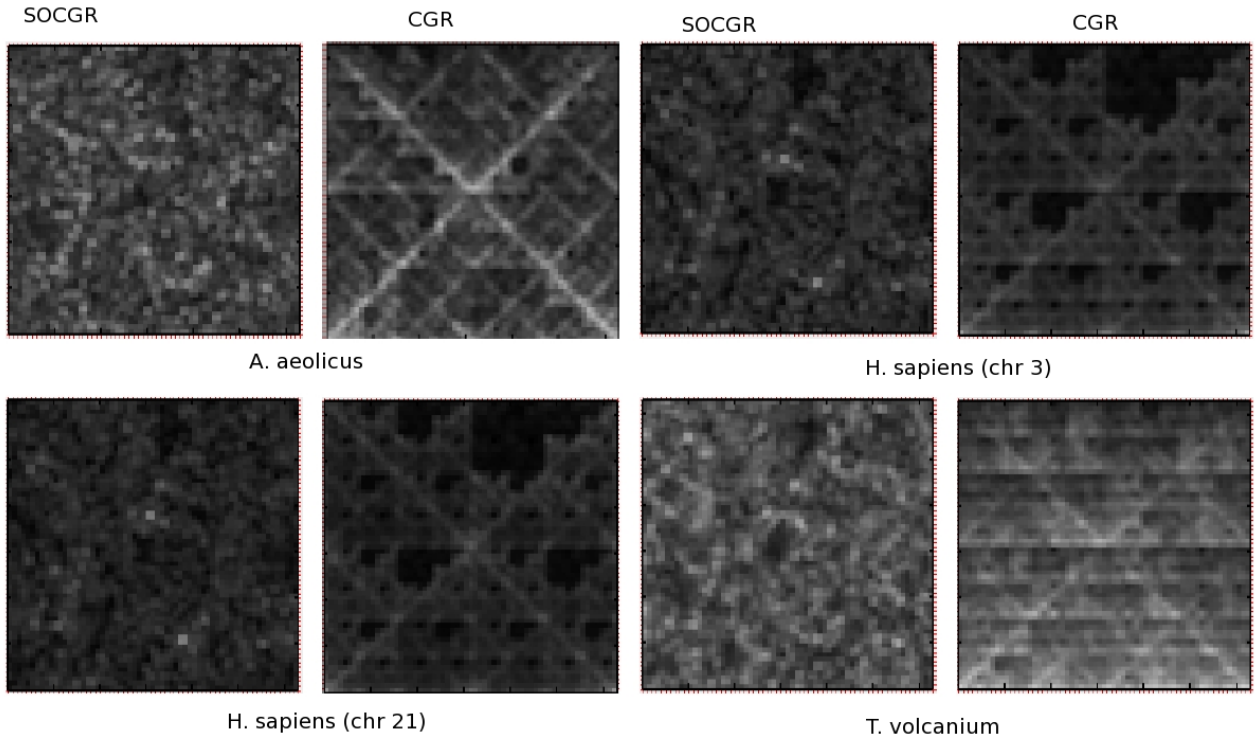

Figure 4. SOCGR (first column) and CGR (second column) for A. *aeolicus*, H. *sapiens* (chromosome 21), H. *Sapiens* (chromosome 3) and T. *volcanium* for sequences of length 6.

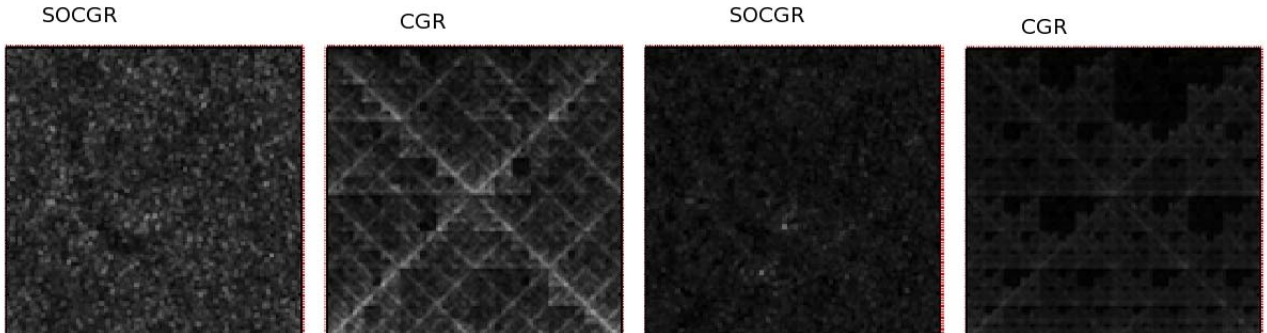

Figure 5. SOCGR (first column) and CGR (second column) for A. *aeolicus*, H. *sapiens* (chromosome 21), H. *Sapiens* (chromosome 3) and T. *volcanium* for sequences of length 7.

Figs. 4 and 5 show SOCGR and CGR for three organisms for the cases of sequence length k = 6 and 7. It is observed that the general structure is maintained in for both levels of detail. Both GS, CGR and SOCGR, are equivalent in the sense that both are permutations of the same space, but the later map presents lower discontinuities and distance between contiguous cells than the former, which represents a better visual GS. Visual inspection is improved if the map presents a lower distance and fewer discontinuities.

One feature of the CGR as a robust genetic signature is that it maintains the overall structure even for low resolution (small values of k). This is observed in figs. 4 and 5, in which the general pattern is conserved regardless the length of sequences. However, the resolution of the pattern is improved as sequence length is increased. These conservation is maintained. These valuable property for GRS is also maintained in SOCGR (see figs. 4 and 5).

As an example, in the SOCGR of *A aeolicus* and *T maritima* , it is observed a pattern in the relative frequencies that is not easy to identify in CGR space, shown in fig. 6. This pattern shows that relative frequency of a group of similar sequences is lower than the relative frequencies of another cluster of sequences. The cells in dark level (low frequency) map for similar sequences and are associated to the sequences TT****G. This pattern is not easy to detect in the CGR space, as it is a discontinuous permutation of cells. Other genomes present a similar pattern such as *T. volcanium*, fig 4, 5.

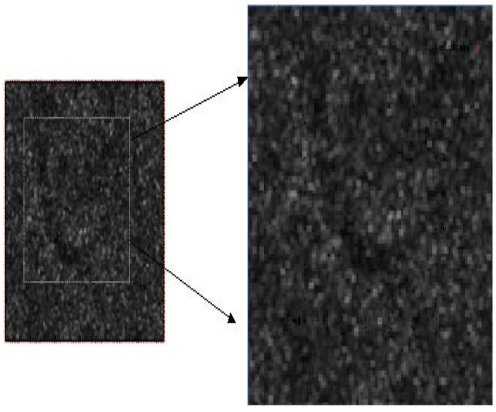

Figure 6. A pattern in the SOCGR of *A aeolicus*. Relative frequency of sequences TT****G is lower than relative frequency of sequences CT****C. In SOCGR contiguous cells tend to be associated to similar sequences, which contrasts with CGR (see figs. 2 and 3)

By analyzing a map in which similar cells or units tend to be associated to similar sequences, as in the case of SOCGR, the discover of relevant patterns in the frequencies of all possible sequences becomes easier. As contiguous cells tend to represent similar sequences, it is natural to seek a local (global) pattern over some sequences by considering the related cell and its neighboring cells.

On the other hand, one of the purposes of the GS is to apply phylogenetic analysis over it, instead of doing it based on the whole genome. Several tools have been applied in order to obtain a phylogenetic structure based on GS, for example, principal component analysis has been used to CGR representations in [5]. SOM has been employed to study di, tri and tetra nucleotide frequencies distribution [10].

Here, we applied the SOM over the SOCGR space in order to seek for hidden features and relationships. It is important to remember that the GS obtained from CGR or SOCGR represents frequency distribution of *k*-lenght sequences. Both, CGR and SOCGR spaces are defined over a $2^k \times 2^k$ lattice, in which each cell represents an unique sequence and all possible *k*-length sequences are represented by a single cell. So, for each cell, it is acquired the relative frequency and it is this value what determines the gray level from figures 4-5. Each genome is represented as a 4^k vector, in which each component is the relative frequency of each sequence. Thus, each genome is represented as a point in the 4^k-dimensional space. Genomes with similar GS are represented by close points in the multidimensional GS space, whereas different genomes are represented by spread apart points.

Fig. 7 shows the map obtained for k = 7 for the organisms listed in table 1. It is observed that related GS from genomes or chromosomes from the same organism are mapped to neighbor neurons, such as organisms 18 and 19 (*H. sapiens* chromosome 3 and *P. troglodytes*). These two organisms are phylogenetically related, and the close relation has been obtained from the analysis of both, the whole genome, or from very conserved sequences such as ribosomal RNA[1]. At the other hand, GS from organisms not phylogenetically related are placed in distant units. The comparison of GS throws results similar to those achieved by comparing complete genomes.

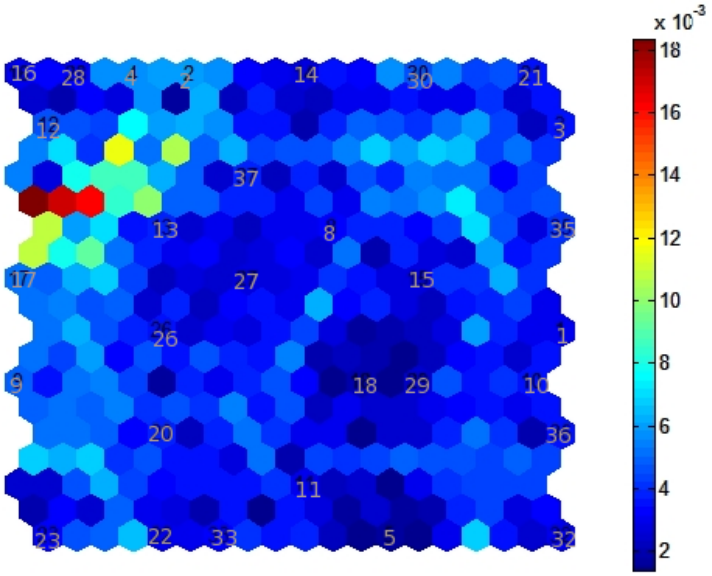

Figure 4. SOCGR (fi nosome 21), H. *Sapiens*

Figure 7. SOM for the SOCGR with sequence *k* = *7*. The numbers correspond to the organisms in table 1.

5. Discussion and Conclusions

Visualization of multidimensional genomic signatures is an important tool to seek for hidden features. Genome analysis is relevant for molecular biologists, as they may infer evolution patterns, phylogenia and many other features. However, as the size of genomes restricts a detailed analysis, several tools have to be applied. Genomic signatures have been proposed as a shorter version of genomes. The Chaos Game Representation has been identified as a good genomic signature. Since it lacks continuity, we propose a variation of it, the Self-Organized Chaos Game Representation. This proposal is based on a a topographic map formed by a self-organizing map, in which similar sequences are mapped to close neurons (cells), which leads to lower discontinuities in the map, opposite to what is observed in the construction of CGR spaces.

The main difference between the SOCGR and CGR is cell distribution. Each cell represents the relative frequency of a given sequence. In the CGR space, very different cells may be placed together, difficulting the exploration process. In the SOCGR space, in which cells that map similar sequences tend to be placed together, image exploration is more intuitive. Pattern discovery is easier when visual analysis is done on a lattice that resembles the distribution of the multidimensional data, as shown in fig. 6. The permutation of cells achieved by SOCGR allows a more intuitive visual exploration, which is a fundamental tool for data mining and pattern-discovering.

As a genomic signature is an abstraction or a summary of a whole genome, it is easier and faster to compare them rather than doing so over the complete genomes. However, if the comparison includes visual inspection, it is a desirable feature that the space analysis preserves neighborhood relationships present in the sequence space. In other words, it is worth that the analysis space to be a topographic map. SOCGR constructs such a space, as similar sequences are mapped to close units whereas dissimilar sequences tend to be mapped to distant units.

6. Bibliographic References

[1] Lewin, R. Genes IX. Jones and Bartlett Publishers. (2007).

[2] Jeffrey, J. Chaos game visualization of sequences. Chaos and fractals: a computer graphical journey. (1988).

[3] Goldman, N. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA. Nucl. Acid. Res. No. 21 (1993) 2487-2491.

[4] Almeida, J., Carriço, J. Maretzek, A., Noble, P., Fletcher, M. Analysis of genomic sequences by chaos game representation. Bioinformatics. Vol. 17 no 5. (2001) 429-437.

[5] Deschavanne, P., Giron, Al., Vilain, J., Fagot, G. Fertil, B. Genomic signatures: characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol. Vol 16 no 10 (1999) 1391-1399.

[6] Karlin, S., Burge, C. Dinucleotide relative abundance extremes: a genomic signature. Trens in genetics. No. 11 (1995) 283-290.

[7] Bauer, M., Schuster, S., Sayood, K. The average mutual information profile as a genomic signature. BMC Bioinformatics. Vol. 9 No. 48 (2008).

[8] Graham, D., Overbeek, R., Olsen, G.,Woese, C. An archaeal genomic signature PNAS Vol. 97 no. 7 (2000) 3304-3308.

[9] Jernigan, R., Baran, R. Pervasive properties of the genomic signatures. BMC Genomics. Vol. 3 No. 23 (2002).

[10] Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T., Ikemura, T. A novel bioinformatic strategy for unveiling hidden genome signatures of eukaryotes: self-organizing map of oligonucleotide frequency. Genome Informatics Vol. 13 (2002) 12-20.

[11] Gatherer, D. Genome signatures, self-organizing maps and higher order phylogenies: a parametric analysis. Evolutionary bioinformatics Vol. 3 (2007) 211-236.

[12] Zhou, F., Olman, V., Xu, Y. Barcodes for genomes and applications.

[13] Carreón, G., Hern´andez, E., Miramontes, P. DNA circular game of chaos. In Statistical physics and beyond. Eds. F. Uribe and L. Garca-Coln. American Institute of Physics. New York. (2005).

[14] Wang, Y., Hill, K., Singh, S., Kari, L. The spectrum of genomic signatures: from dinucleotides to chaos game representations Gene. No. 346 (2005) 173-185.

[15] Kohonen, T. Self-Organizing maps. 3rd. ed. Springer-Verlag. (2000).

[16] Cottrell, M. Fort, J.C., Pag´es, G. Theoretical aspects of the SOM algorithm. Neurocomputing 21 (1998) 119-138.

[17] Ritter, H. Self-Organizing Maps on non-euclidean Spaces Kohonen Maps, 97-108, Eds.: E. Oja and S. Kaski, (1999).

[18] Neme, A., Miramontes, P. A parameter in the SOM learning rule that incorporates activation frequency. ICANN (1) (2006) 455-463.

[19] Benson, D., Karsch-MizrachiD., Wheeler, D. GenBank. Nucleic Acids Res. 36 (2008). D25-30.

[20] Chang, C., Xu, P. Frequency sensitive self-organizing maps and its application in color quantization. Proc. Of the Int. Symp. In Circuits and Systems Vol. 5(2004), 804-807.