DNA Circular Game of Chaos

Gustavo Carreón, Enrique Hernández[*] and Pedro Miramontes[†]

[*]*Program of Nonlinear Dynamics and Complex Systems*
Universidad Autónoma de la Ciudad de México, México 03200 DF, México
[†] *Department of Mathematics, Faculty of Sciences*
Universidad Nacional Autónoma de México
Cd. Universitaria, México 04510 DF, México.

Abstract.
One of the most important aims in evolutionary biology is the search of historical as well as structural relationships among species. In this report, we show that traditional and well-known results from the theory of nonlinear dynamics can provide a useful ground to achieve this end. In particular, we propose whole genome phylogenies treating DNA as a discrete sequence and then feeding it to a dynamical system.

1. THE CIRCULAR GAME OF CHAOS

Two decades ago, an algorithm first described by M. F. Barnsley [2] was named "The Game of Chaos". Originally the Game was meant to be played on a triangle with the following rules:

1. Select in the coordinate plane three non-collinear points to be the vertices of the triangle.
2. Randomly select one point inside the triangle and call its coordinates (x,y).
3. Randomly select a vertex and draw a point halfway between (x,y) and the vertex.
4. Redefine (x,y) to be the coordinates of the new point and repeat the previous step

This algorithm generates the well known Sierpinski triangle. When instead of a triangle, a square is used to play the Game, a random four-symbol sequence assigned to the vertices as in step 3, fills the area densely (a hypothetical perfect random sequence fills out the square uniformly). The game of chaos on a square can tell about the properties of any four-symbol sequence. As a matter of fact, it is an important tool to tell apart diffrent types of noise (white noise, Brownian motion, chaos or flicker noise [9]).

In 1999, Tsuchiya Takashi [13] realized that further insights about a dynamically generated discrete sequence could be obtained by increasing the number of sides of the polygon and getting in the limit what he named The Circular Game of Chaos (it is important, since the very outset, to remark that the denomination *game of chaos* is a misnomer because the relationship of the fractal structures first obtained by Barnsley on a triangle to what the community understands by deterministic chaos is tangential).

And interesting scenario opens out when the Circular Game of Chaos (CGCh) is played with a binary sequence over a *n*-sided polygon (*n* being a power of two), then each vertex is labeled with a binary word of length *n*. If the Game is played by sliding,

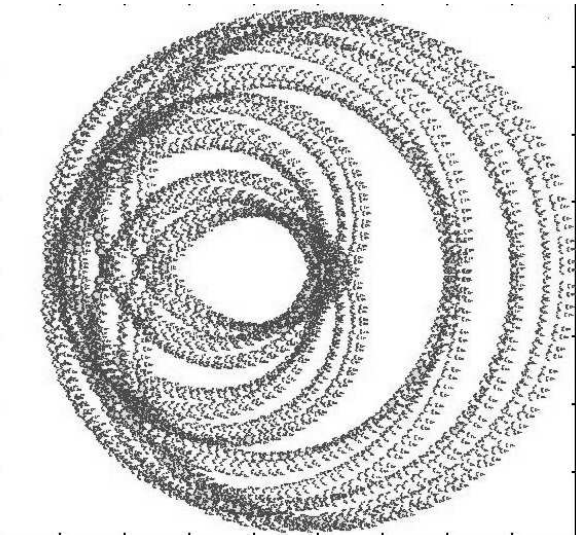

FIGURE 1. Output of the circular game of chaos over a windowed binary sequence

one step at a time, an *n*-length window over a binary sequence, then the figure that appears is the fractal shown in Figure 1.

The fractal in Figure 1 is relatively independent of the binary sequence chosen as long as it is no constant or periodic. As a matter of fact, it mainly reflects the characteristics of the shift operator over binary sequences but it is expected to show different density of iterates depending upon on the nature of the feeding sequence (random, colored noise, chaotic, etcetera).

2. DNA CIRCULAR GAME OF CHAOS

In a first approximation, DNA can be studied mathematically as a sequence of four symbols A, C, G and T (corresponding to the four types of residues -or bases- Adenyl, Cytidyl, Guanyl and Thymidyl). The whole sequence constitutes an organism's *genome*, the length of the genome from different organisms ranges from 10^3 bases in some viruses to 10^{12} in some species of salamanders. It became very soon evident that a DNA sequence could feed a Game of Chaos on a square [5] by labeling each vertex with one base and getting results similar to those of Figure 2.

The fractal structure of DNA was reported almost as soon as genomic databases were available [12]. Initially, it was disclosed when several researchers treated the molecule as a four-symbol abstract sequence and used spectral and time-series analysis to obtain slow decaying correlograms and spectra. Before that, molecular biologists already knew that chromosome banding shows self-similarity at several magnitude scales [1]. Further

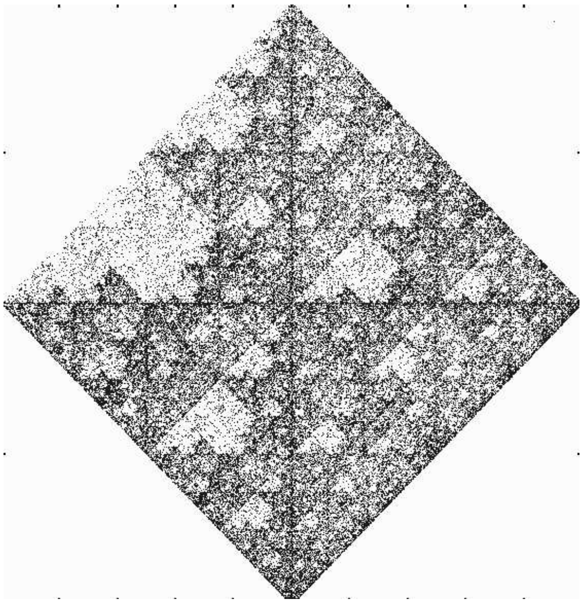

FIGURE 2. The DNA Game of Chaos of the bacterium Agrobacterium-tumefaciens over a square. Notice that the fractal-like structure is a reflect of the structure of correlations of the genome mapped

investigations resulted in the existence of DNA power-law scaling and long-range correlations [11] [7]. Several models were proposed to explain these facts [6].

Any DNA sequence can be translated into three different binary sequences according to the way pair groups are formed. The first grouping $((A,T) \to 0,(C,G) \to 1)$ is be called *WS dichotomy*. The YR and MK dichotomies [1] are defined analogously: $((A,G) \to 0,(C,T) \to 1)$ and $(A,C) \to 0,(T,G) \to 1)$.

Any of these binary sequences can be used to play the DNA Circular Game of Chaos (DNACGCh)

Figures 3 and 4 are the result of a DNACGCh using the WS DNA dichotomy. They are 2-d density normalized histograms of the already mentioned shift attractor.

The Figures correspond to two species of Bacterial. It is evident that, while having the same support set on the plane, the histograms are pretty different for different organisms even if they are phylogeneticaly related. In this research we calculated the histograms of a couple of dozens of organisms representing the three known domains (Eucarya, Bacteria and Archaea).

[1] WS,YR,MK stand for Weak-Strong, pYrimidine-puRine and aMino-Ketone, these dichotomies are closely related to structural, chemical and thermodynamical properties of the DNA molecule. See [8]

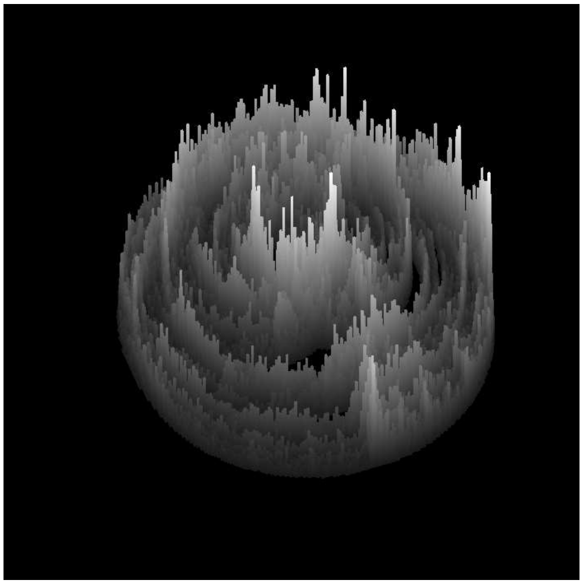

FIGURE 3. Histogram of iterates density for the shift operator using the *Thermotoga maritima* genome

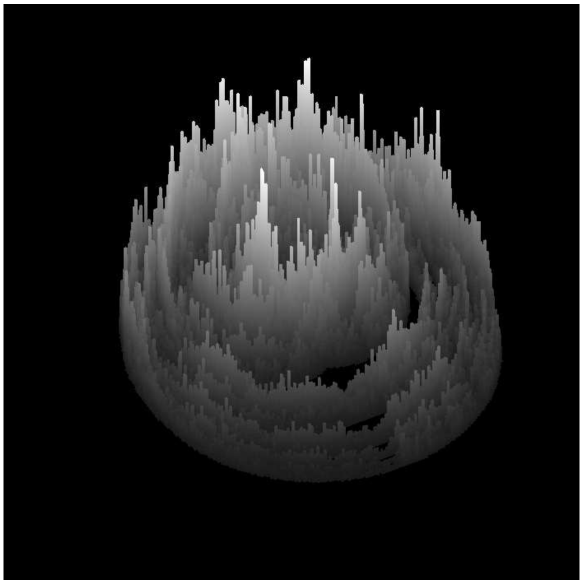

FIGURE 4. Histogram of iterates density for the shift operator for the *Bacillus subtilis* genome

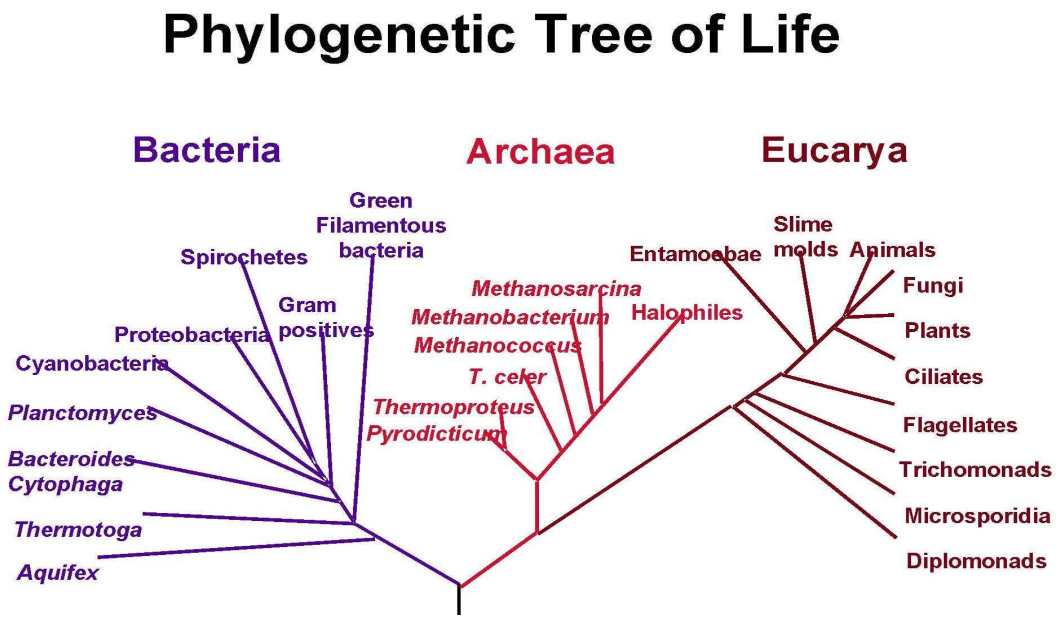

FIGURE 5. The phylogenetic “tree of life"

3. GENOME PHYLOGENIES

One of the aims of Biology is to find out historical, functional and structural relationships among organisms and to use this information to group them into families or categories depending on the nature of the traits used to carry out the grouping. The result is a phylogenetic tree like the one showed in Figure 5. Speaking in mathematical jargon; the task is to assign to each species (or the proper taxon) a vector of traits or characteristics, then to calculate a matrix of distances and finally to group the vectors (organisms) using a standard cluster analysis to obtain a tree or dendrogram.

No doubt that the results can be dramatically different depending on the way the vector of traits is associated to each organism. Traditional taxonomy used to employ morphological traits to classify organisms but with the advent of genomic databases the construction of dendrograms based upon the distance between genes has become a standard procedure.

Notwithstanding its general acceptance, the above mentioned method is anything but free of problems. There are a lot of technical details outside the scope of this report that can be consulted in the book of J. Felsenstein [4].

4. FRACTAL PHYLOGENIES

The normal procedure to construct a tree out of genomic data begins with the process of DNA alignment. It is the adjustment of the position of two or more molecular sequences relative to each other so that similar positions of the molecule can be put together. A pair of DNA sequences (a pair of four-symbol sequences) are written in consecutive rows and one of them is slided until the statistical resemblance between the two is maximum. It is allowed to introduce gaps to increase the resemblance, and it is allowed to put in the same position different symbols. In the Following example:

$$\begin{array}{ccccccccccc} A & C & C & G & - & T & T & A & C & G & G \\ A & C & A & G & C & T & - & - & C & G & G \end{array}$$

The distance between the sequences will be inversely proportional to the amount of

coinciding symbols minus a punishment for the forced introduction of gaps and different symbols in the same position.

Multiple alignment techniques can be very sophisticated [4] but the basics lay on the idea of pairwise alignment. Once in possesion of of a matrix of distances, common cluster analysis can be performed in order to get a tree of similarities.

One of the yet unsolved problems in molecular phylogeny is that genes are small parts of a genome (the characteristic size of a gene ranges between hundreds to thousands of bases) and that the choice of different genes can lead to results where the grouping is different.

Our proposal is to make whole genome phylogenies independent of the protein coding constraints of DNA and based solely in the structural properties of the molecule. The main problem is then how to associate vector of the same length to genomes that can be different by many orders of magnitude. To accomplish this, we recurred to fractal image compression techniques.

Developed by Barnsley [3] the fractal compression idea is awesome; to compress an image, just develop a dynamical system in such a way that the original image is the attractor of this dynamical system. Given a random initial condition (a point in the plane) the continued iterations of the dynamical systems would eventually converge to the image.

The details of the fractal compression algorithm can be found in [3] but the idea is simple: Consider a hybrid dynamical system (deterministic rules probabilistically applied):

$$x(t+1) = \begin{cases} F_1(x(t)) & \text{with } p = p_1; \\ F_2(x(t)) & \text{with } p = p_2; \\ \vdots & \\ F_m(x(t)) & \text{with } p = p_m \end{cases}$$

Where t is time in discrete steps, $x \in \mathbb{R}^2$, F is a contractive function, usually an affine transformation, and p is the probability distribution with which every F_i is applied.

We applied the fractal image compression algorithm to the circular game of chaos graphical output corresponding to four Bacteria, three Eucarya (*H. sapiens*, a worm and a plant) and three Archaea.

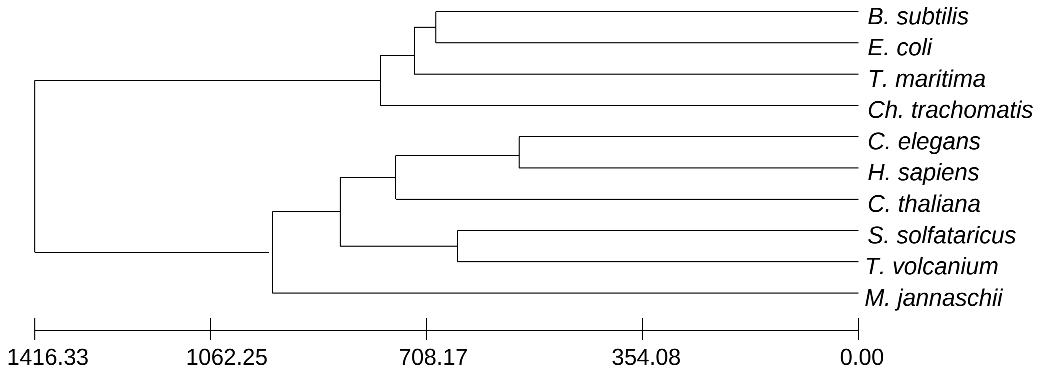

FIGURE 6. An example of whole genome dendrogram. The horizontal units of similarity are arbitrary

5. RESULTS AND DISCUSSION

Figure 6 is the summary of the research described in the above paragraphs. It is a dendrogram relating ten organisms having quite contrasting genome sizes. The process, in a concise way, was:

1. The genomes where translated to the *WS* binary representation. This dichotomy was chosen over the remaining two because it has a strong phenomenological meaning; it is related to the stability of the DNA double-helical structure. See [8].
2. A Circular Game of Chaos scenario was arranged: a 512-sided polygon was set and each vertex was labeled with a 9-long binary word.
3. A moving 9-long window was slided, one step at a time, over the binary sequence and the step 3 in the game of chaos was carried out.
4. The resulting fractal like image (similar, but not equal to Figure 1) was normalized and compressed using the Fractal Compression Algorithm and in this way obtaining a vector associated to every species.
5. A nearest-neighbor pairwise cluster analysis was performed and the result is as shown in Figure 6.

The adoption of this procedure by the biological community is something still to be seen. But the potential is not to be disdained: 1. The bacteria *T. maritima* was hard to classify because it has many Archaea characteristics. It even has a high amount of Archaeal DNA acquired by horizontal gene transfer [10]. Our procedure puts it deeply into the Bacterial branches of the tree. 2. *Ch. trachomatis* is a intracellular parasite. It is bacteria with no free life outside an Eukaryan cell. It has been suggested [8] that the forced coevolution with Eucarya has pushed *Ch. trachomatis* to adopt several structural DNA traits of its guests. In our tree, this organism is not far from Eukarya. 3. The fact that Archaea is closer to Eucarya than to Bacteria is an amazing but well established fact [14]. 3. *M. jannaschii* is an odd organism: It was first isolated from a sediment coming from a 2600-m-deep "white smoker" chimney located on the East Pacific rise. It is an

Archea but our method places it closer to Eucarya. The biological interpretation of this fact is still to be found.

The final comment is that well-established facts from Nonlinear Dynamics could still be worthwhile. In this case we showed that the Iterated Function Systems proposed by Barnsley back in the eighties still have spring to unwind.

REFERENCES

1. Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M and F. Rodier. *Science*, **228**, 1985.
2. Barnsley Michael F., *Fractals Everywhere*, Academic Press, 1993.
3. Barnsley Michael F., *Fractal Image Compression*, Ak Peters Ltd., New York, 1993.
4. Felsenstein J. *Inferring Phylogenies*, Sinauer Associated. 2003.
5. Jeffrey HJ. *Nucleic Acids Res*, **18**, 1990.
6. Li W. *Computer & Chemistry* **21** (1997)
7. Li W. and K. Kaneko, *Europhys Lett* **17** (1992) pp. 655-660.
8. Miramontes P., Medrano L., Cerpa C., Cedergren R., Ferbeyre G. and G. Cocho, *J Mol Evol* **40**, 1995.
9. Miramontes O. and P. Rohani. *Proc R Soc Lond B Biol Sci* **265**, 1998.
10. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, Fraser CM, et al. *Nature*, **399**, 1999
11. Stanley HE, Buldyrev SV, Goldberger AL, Hausdorff JM, Havlin S, Mietus J, Peng C-K, Sciortino F and M. Simons *Physica **A***, **191**, 1992
12. Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M and HE Stanley. *Nature*, **356**, 1992
13. Tsuchiya T. *Internat J Bifur Chaos*, **9**, 1999
14. Woese CR, Fox GE. *Proc Natl Acad Sci USA*, **11**, 1977