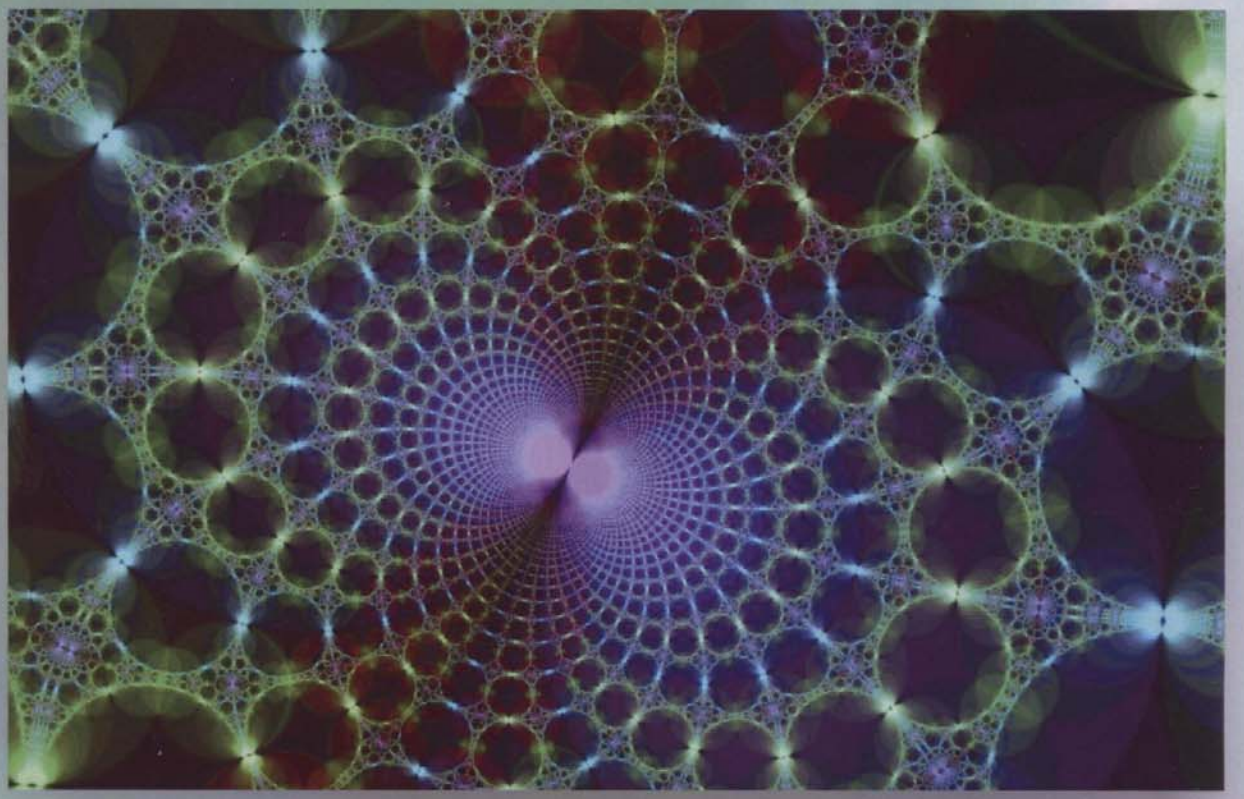
DYNAMICAL SYSTEMS

— Applications —



EDITORS

Jan Awrejcewicz, Marek Kaźmierczak

Paweł Olejnik, Jerzy Mrozowski

Fractional analysis of the DNA information

José António Tenreiro Machado, Alexandra Maria S. Ferreira Galhano

Abstract: This study addresses the comparison of the deoxyribonucleic acid (DNA) of six primates in the perspective of fractals and signal processing. The species consist of Bonobo, Chimpanzee, Gorilla, Human, Orangutan, and Rhesus macaque. These primates are very close in the life evolution and, therefore, pose a problem for designing assertive algorithms to detect their phylogenetic tree. The paper associates logical and mathematical concepts inspired in signal processing and dynamical systems for the analysis of the DNA data of the chromosomes. The results are compared by means of computer visualization tools.

1. Introduction

Phylogenetics studies the evolutionary relations between groups of organisms. The progresses with genome sequencing and genome databases led to the emergence of considerable data presently available for computational processing. The informational structure embedded into the deoxyribonucleic acid (DNA) can be exploited for constructing more precise phylogenetic relationships. Understanding the DNA and extracting information is a challenging problem that motivates worldwide research (Pearson, 2006, Seitz, 2007). The paper proposes an algorithm for converting the DNA string into a numerical signal. Once established this quantifying procedure, the concepts of state space, signal, fractal and computer visualization tools are adopted.

Are addressed six primates, namely the Bonobo, Chimpanzee, Gorilla, Human, Oranguta and Rhesus macaque, constituting species that are very close in evolutionary terms. Therefore, the corresponding set of chromosomes (chr) poses a formidable challenge for distinguishing details and extracting assertive phylogenetic information.

Having these ideas in mind, this paper is organized as follows. Section 2 presents the main genetic details of the primates under studied and proposes the scheme for translating DNA information into a signal. Sections 3 and 4 develop the DNA sequence processing and phylogenetic analysis based on fractal and dynamical analysis, respectively. Finally, section 5 outlines the main conclusions.

2. Fundamental concepts

DNA consists of a double helix with a sequence of four nitrogenous bases {Thymine, Cytosine, Adenine, Guanine} represented by the symbols $\{T, C, A, G\}$. The chr data files includes a fifth symbol N which is considered to have no practical meaning for the DNA decoding. Usually the percentage of N is relatively smaller than the rest of the symbols. Each base present in one side connects only with another type of base on the second side of the double helix and forms the base pairing AT and CG. This information is being collected and is available for scientific research (Murphy, 2007,Ebersberger, 2007,Dunn, 2008, Prasad, 2008, Sims, 2009, Zhao, 2009). DNA embeds an intricate information, seemingly with distinct scales and redundancies. Therefore, mathematical tools adopted in the study of dynamical systems may lead to an assertive analysis and quantitative results (Machado, 2011).

In the sequel we consider 6 primates {Bonobo, Chimpanzee, Gorilla, Human, Orangutan, Rhesus macaque} denoted as {Bo, Ch, Go, Hu, Or, Rh} making a total of 143 chrs.

The first phase of the proposed algorithm consists of translating the string of symbols into a string of numbers. The conversion of the DNA 4(+1)-symbol into a numerical signal is accomplished with a 2(+1)-dimensional space representation. First the AC and TG pairs are represented in the Cartesian coordinates x and y, respectively. The position along the DNA length L is represented by means of the z Cartesian axis. Second, each successive symbol in the DNA is converted to a one-step increment being $+1$ (-1) for the first (second) base in each bonding pair. For the symbol N no action is taken. By other words, our algorithm maps the genome sequence onto the points $P_k = (x_k, y_k, z_k)$ of a 3-dimensional random walk used the following 2 rules:

1. The walk starts at $P_0 = (0, 0, 0)$;

2. The nucleotides of a genome are read in succession, the coordinates of the next point $P_{k+1} = (x_{k+1}, y_{k+1}, z_{k+1})$ are obtained from present point $P_k = (x_k, y_k, z_k)$ as:

$$x_{k+1}, y_{k+1} = \begin{cases} x_k + 1, y_k + 1 & \text{if the next nucleotide is A,C} \\ x_k - 1, y_k - 1 & \text{if the next nucleotide is T,G} \\ x_k, y_k & \text{otherwise} \end{cases}, \quad z_{k+1} = z_k + 1$$

This algorithm preserves the base pairing logic and does not introduce any preconception biasing the DNA information. The 3- and 2- (i.e., the projection over the horizontal plane) dimensional representations are going to be considered. These spaces will be denoted as the locus $\{AT, CG, L\}$ and $\{AT, CG\}$, respectively. For example, Figure 1 shows the trajectory

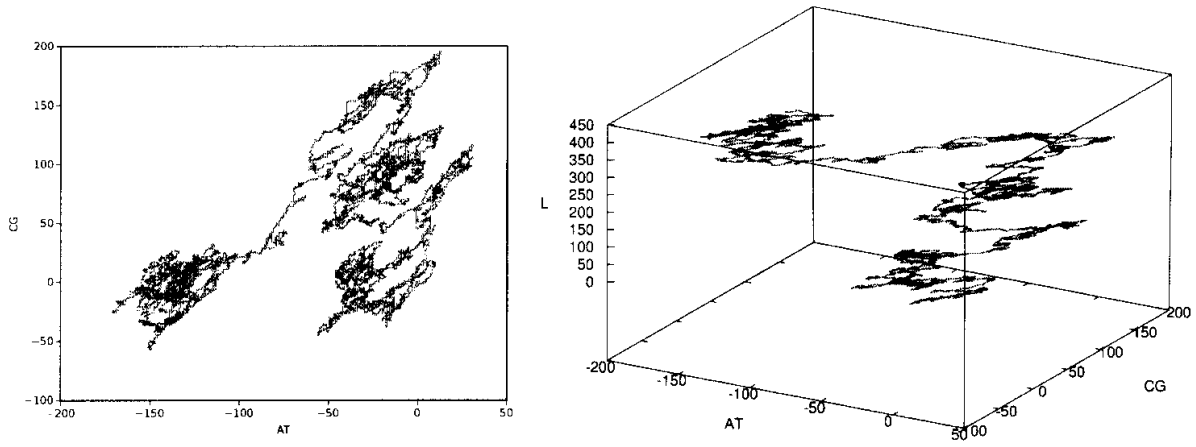


Figure 1. Trajectory locus $\{AT, CG\}$ and $\{AT, CG, L\}$ of the chr 1 of the Bonobo.

$\{AT, CG\}$ and the corresponding locus $\{AT, CG, L\}$ for the chr 1 of the Bonobo and a quantifying cube of $500 \times 500 \times 500$.

According with the second Chargaff's rule the number of symbols $\{T, C, A, G\}$ is approximately identical, not only for each of the two DNA strands, but also for long sequences (Mitchell, 2006). Nevertheless, in the present case we are capturing the *order* of the symbols along the sequence and, therefore, considerable deviations from the 45 degree line in the x, y projection plane may occur. Computation of the complexity for DNA representations is interesting and we can mention the Z-curve (Zhang, 2003). Furthermore, it is worth nothing the so-called C-value paradox which states that organism complexity does not correlate with genome size.

3. Fractal analysis of the DNA

The second phase consists of extracting information from the 3-dimensional trajectories. Since the plots reveal close resemblances with fractals it is adopted the box-counting method for their measurement (Berry, 1979, Lapidus, 1988, Schroeder, 1991). For an object S lying in a n-dimensional space, let $N_\varepsilon(S)$ be the minimum number of n-dimensional cubes of side-length $\varepsilon > 0$ needed to cover S. If there is a number b so that $N_\varepsilon(S) \sim \frac{1}{\varepsilon^b}$ as $\varepsilon \to 0$ we say that the box-counting dimension of S is b. This leads to the expression:

$$N_\varepsilon(S) = a\frac{1}{\varepsilon^b} \tag{1}$$

where a and b are the parameters that captures the size and the dimension of the object S. In our case S consists of the 3-dimensional DNA representation in the locus $\{AT, CG, L\}$.

The six primates involve to a total of 143 chrs to be analysed. Each chr is converted into a representation in the locus $\{AT, CG, L\}$. In a first step the chr information is read and

the maximum and minimum limits along the (x, y, z) axes is determined. This information allows the calculation of a common scale factor so that the 3-dimensional trajectories have dimensions that reflect the information content and the chr length L. Once calculated the maximum/minimum limits and the global scale factor, each chr is read again and the corresponding trajectory in the locus $\{AT, CG, L\}$ is plotted. Finally, the resulting fractal trajectory is measured by means of (1) and the parameters (a, b) extracted.

Figure 2 shows the locus (a, b) for the 143 chrs. The parameter a reflects the length L of the chr. Therefore, we observe a tendency for smaller/larger values of the point labels in the right/left of the locus of (a, b). The parameter b is related with the information content, being lower/higher as the DNA "signal" has a fractal structure that resembles more a common line/surface. There is also a separation in the perspective of the six primates. It is visible Rm at the bottom and Go at the left, clearly apart from the rest. The group {Bo, Ch, Ho, Or} is very close and becomes more difficult to distinguish. There are differences from chr to chr but, in general, it is difficult to establish some order between them.

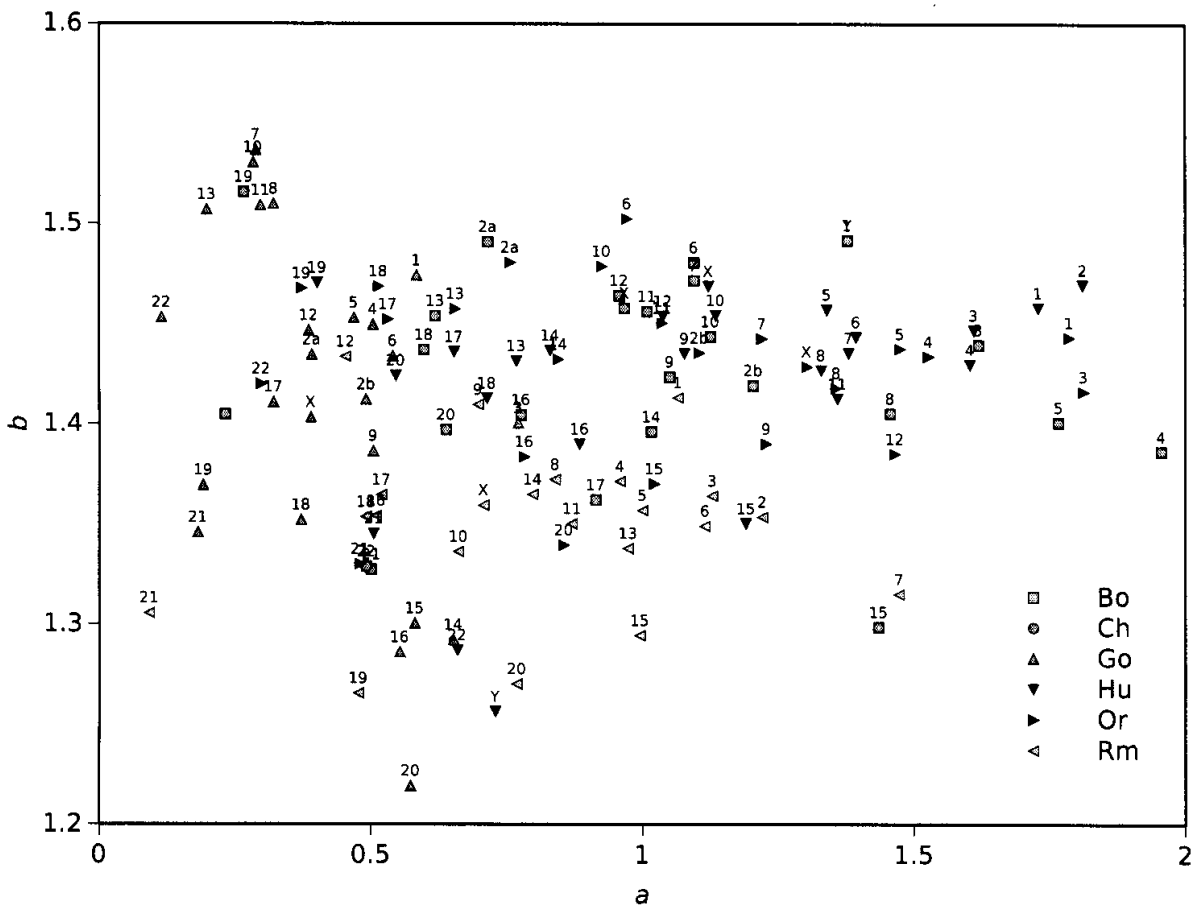


Figure 2. Locus (a, b) of the 143 chromosomes.

In the numerical representation $\{AT, CG, L\}$ the L coordinate may not provide additional scale information of the DNA sequence. By other words, the $\{AT, CG\}$ coordinates are those of the 2-dimensional DNA walk model of DNA sequences (Luo, 1998). Fractal methods have been used to construct the phylogenetic trees (Yu, 1998, Yu, 2003, Yu, 2004). It has been discussed that the fractal methods cannot give better phylogenetic trees than simple statistical models if the number of species becomes large. Therefore, the adoption of the fractal dimension for characterizing the objects constructed in the $\{AT, CG, L\}$ locus seems to limit the visualization and that a more suitable method should be tried.

4. Dynamical analysis of the DNA

In this section it is adopted a method inspired in dynamical systems analysis. The main idea comes from the fact that the z information can be considered as an "indexing" of the signal in the $\{AT, CG\}$ 2-dimensional space. Then, the signals can be compared among themselves and the results depicted by means of a suitable computer visualization tool.

The signals $\{x_i(z), y_i(z)\}$, $z_i = 0, \cdots, L_i$, and $\{x_j(z), y_j(z)\}$, $z_j = 0, \cdots, L_j$, of the i-th and j-th chrs, are compared by means of the 2-dim correlation measure (Deza, 2009):

$$r_{ij} = \left| \frac{\sum_{z=0}^{min(L_i, L_j)} \left[x_i(z) x_j(z) + y_i(z) y_j(z) \right]}{\sqrt{\sum_{z=0}^{L_i} \left[x_i^2(z) + y_i^2(z) \right] \sum_{z=0}^{L_j} \left[x_j^2(z) + y_j^2(z) \right]}} \right| \tag{2}$$

where $min(L_i, L_j)$ denotes the minimum of the lengths, L_i and L_j, of the two signals. Since the chrs have distinct sizes, in (2) it is considered that the shortest signal has value zero after exceeding its length.

The index r_{ij} runs over the set of chrs performing an item-to-item comparison and producing a symmetrical matrix $R = [r_{ij}]$ that can be analysed using present day visualization tools. For comparing the chrs of the 6 primates it is produced a matrix R, 143×143 dimensional. For the visualization it is adopted the package PHYLIP (Tuimala, 2006) with the Fitch-Margoliash algorithm and the "drawgram" and "drawtree" plotting methods, that are presented in Figures 3 and Figures 4. We observe again that the Ch and Bo are the species closer to the Hu, while the Go and Rm are considerably different.

Besides the 2-dimensional visualization program adopted previously, it is also tested the Multidimensional scaling (MDS) by means of the algorithm 'ggvis" in the package GGobi (Buja, 2004, Lang, 2006). MDS is a statistical technique for visualization information that explores similarities/distances in the data (Shepard, 1962, Kruskal, 1964, Cox, 2001, Borg, 2005). It arranges items in a space with a given number of dimensions, so as to reproduce

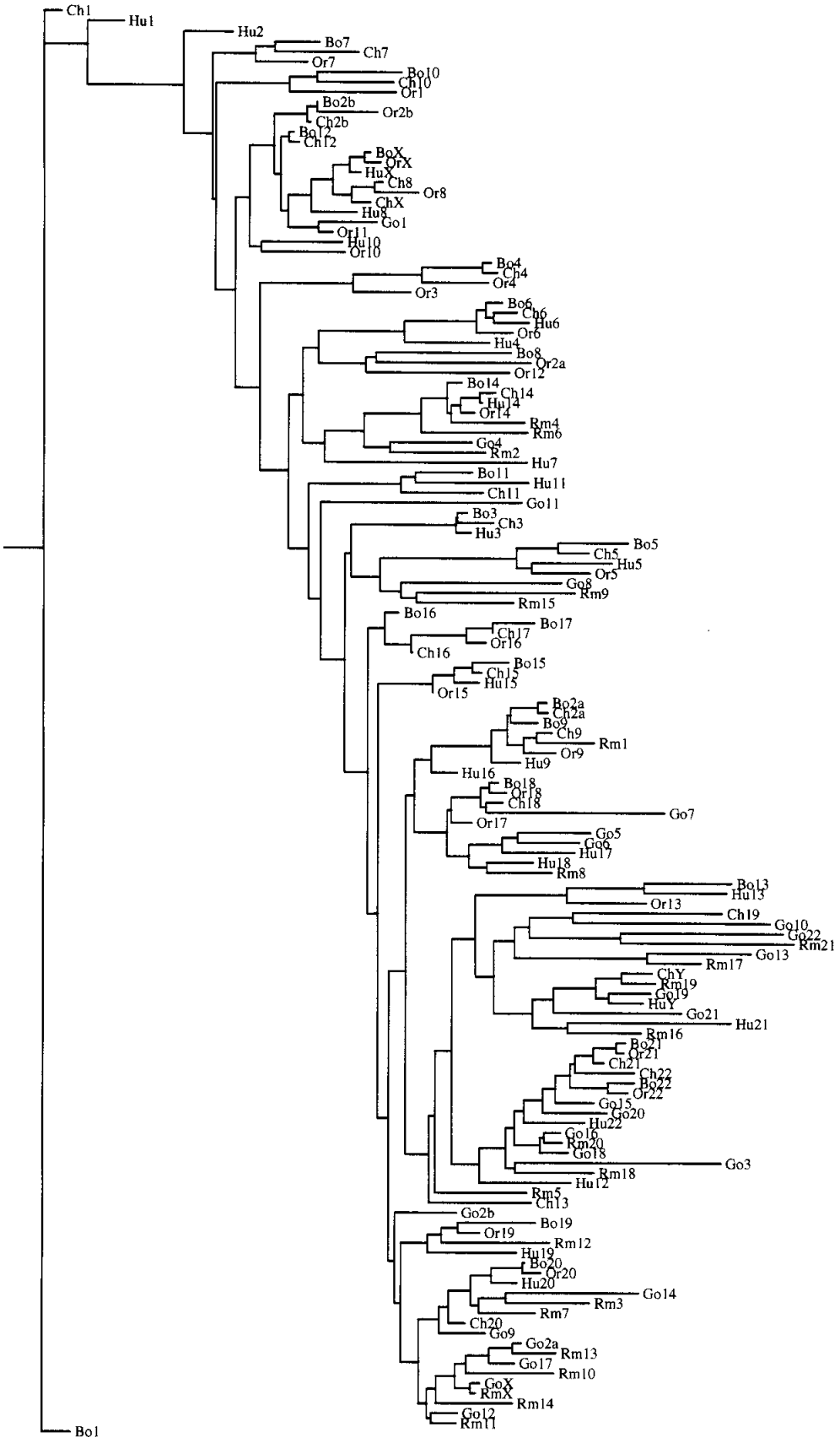


Figure 3. Comparison of the chrs of the 6 primates using (2) and visualizing by means of a dendogram.

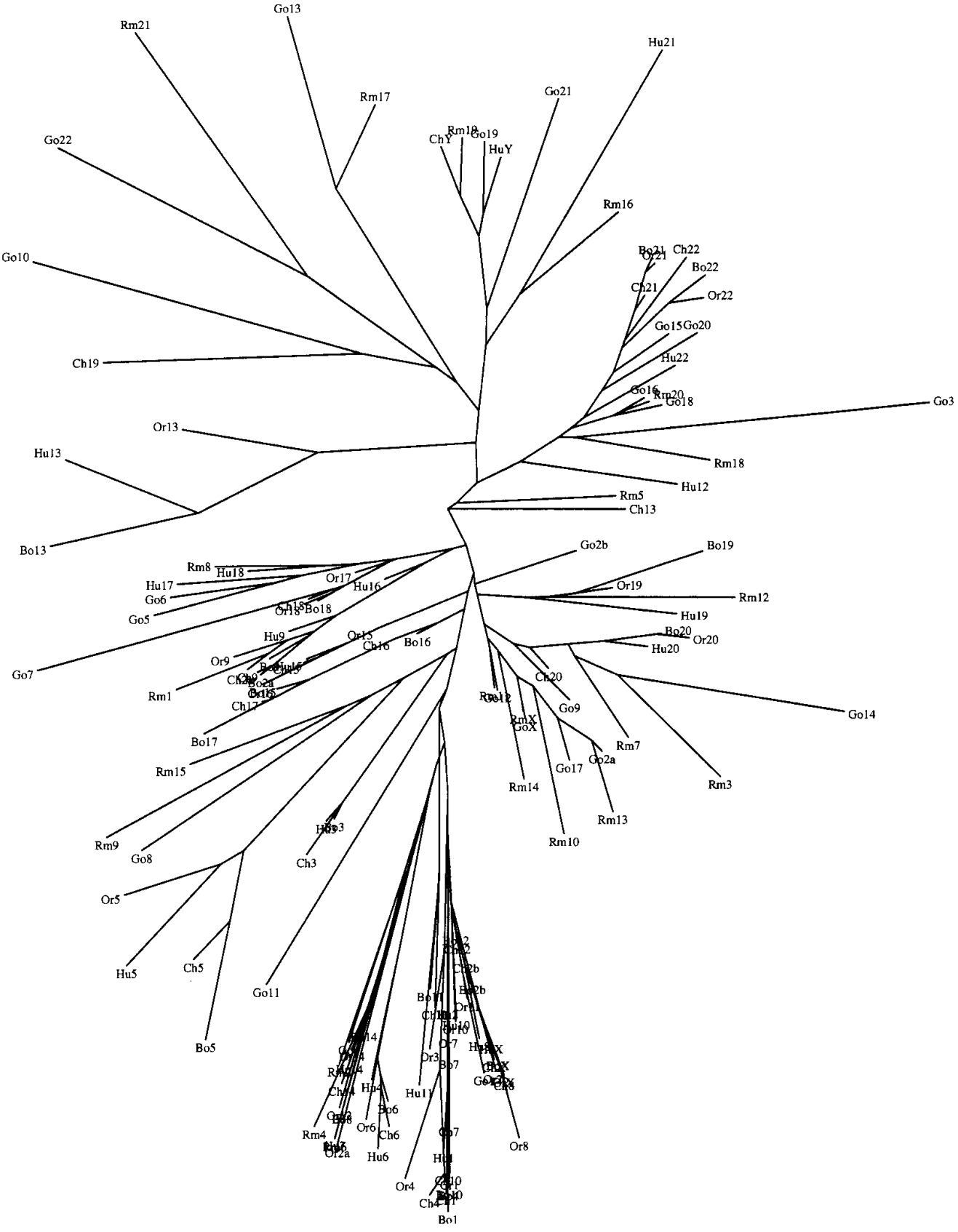


Figure 4. Comparison of the chrs of the 6 primates using (2) and visualizing by means of a 2-dimensional tree.

the observed similarities between the group of objects. Figure 5 shows the 3-dimensional MDS plot for the 6 species based on index (2).

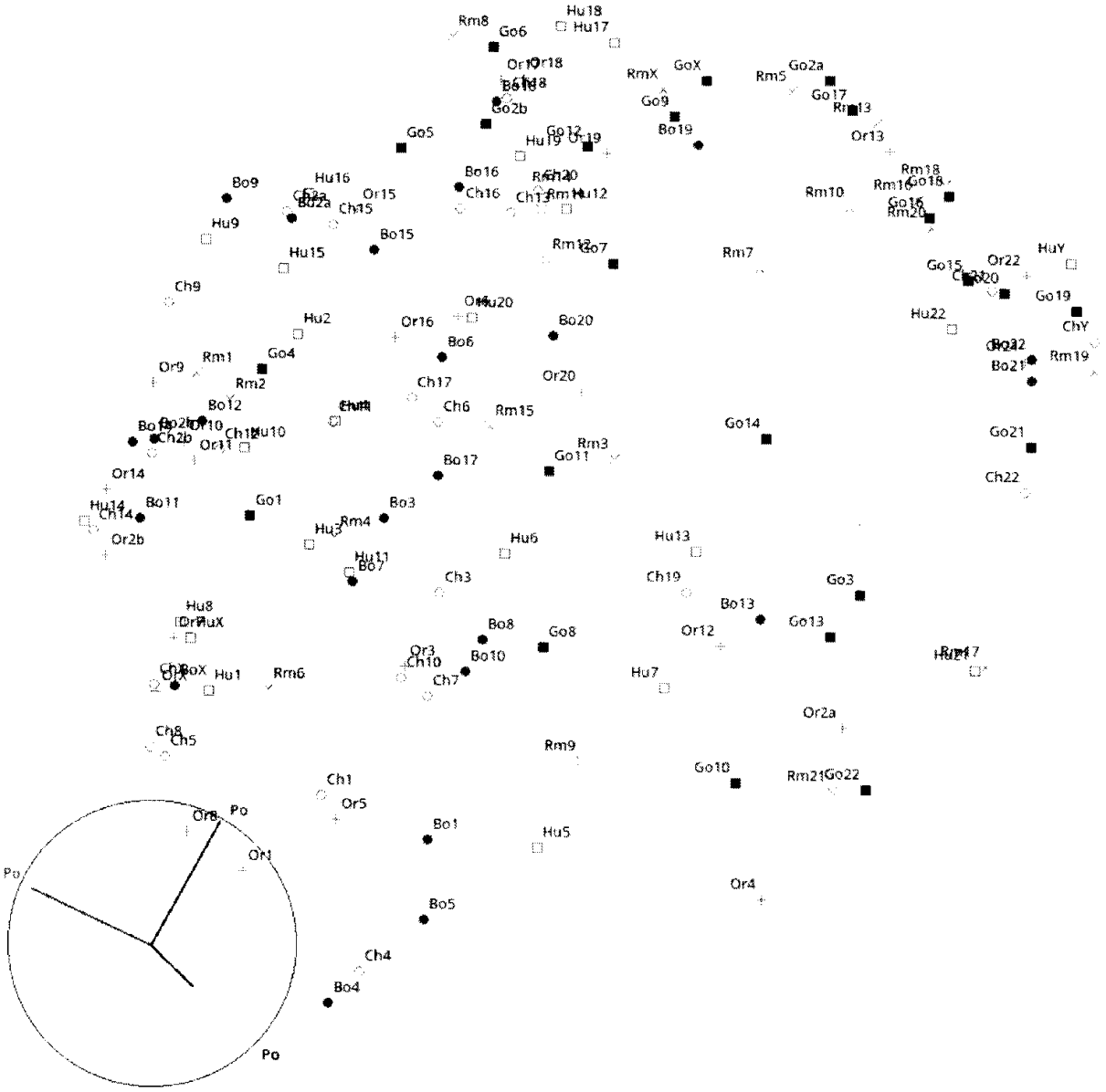


Figure 5. Comparison of the chrs of the 6 primates using (2) and visualizing by means of a 3-dimensional MDS plot.

We observe again that the Ch and Bo are the species closer to the Hu, while the Go and Rm are considerably different. Since the set {Bo, Ch, Hu} are the closest set we decided to test the proposed methodology with the three species.

5. Conclusions

In this paper was analysed the DNA of the Bonobo, Chimpanzee, Gorilla, Human, Orangutan and Rhesus macaque. These primates are very close and pose considerable difficulties for distinguishing them in the evolutionary tree. It was proposed an algorithm for converting the DNA information into a numerical "signal". The resulting object revealed a fractal structure but its description by geometrical dimension leads to limited results. It was also developed an approach inspired in signal processing by defining the cosine correlation measure. The comparison of the chrs was then performed and the results visualized taking advantage of modern computer tools.

References

1. Berry M.: Diffractals, *Journal of Physics A: Mathematical and General*, 12(6), 1979, 781-797.
2. Borg I., Groenen P.: *Modern Multidimensional Scaling-Theory and Applications*, New York, Springer-Verlag, 2005.
3. Buja A., Swayne D.F., Littman M.L., Dean N., Hofmann H.: *Interactive data visualization with multidimensional scaling*, 2004.
4. Cox T., Cox M.: *Multidimensional Scaling*, Boca Raton, Chapman & Hall/CRC, 2001.
5. Deza M., Deza E.: *Encyclopedia of Distances*, Berlin, Springer-Verlag, 2009.
6. Dunn C.D. et al.: Broad phylogenomic sampling improves resolution of the animal tree of life, *Nature*, 452(10), 2008, 745-750.
7. Ebersberger I., Galgoczy P., Taudien S., Taenzer S., Platzer M., Haeseler A.: Mapping human genetic ancestry, *Molecular Biology and Evolution*, 24(10), 2007, 2266-2276.
8. Kruskal J.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, 29(1), 1964, 1-27.
9. Lang D.T., Swayne D.F.: *The GGobi XML input format*, 2006.
10. Lapidus M., Fleckinger-Pellé J.: Tambour fractal: vers une résolution de la conjecture de Weyl-Berry pour les valeurs propres du laplacien, *Comptes Rendus de l'Académie des Sciences Paris Sér. I Math.*, 306, 1988, 171-175.
11. Luo L., Lee W., Jia L., Ji F., Tsai L.: Statistical correlation of nucleotides in a DNA sequence, *Physical Review E*, 58(1), 1998, 861-871.
12. Machado J.T., Costa A., Quelhas M.: Entropy analysis of DNA code dynamics in human chromosomes, *Computers and Mathematics with Applications*, 62(3), 2011, 1612-1617.
13. Machado J.T., Costa A., Quelhas M.: Shannon, Rényi and Tsallis entropy analysis of DNA using phase plane, *Nonlinear Analysis Series B: Real World Applications*, 12(6), 2011, 3135-3144.
14. Mitchell D., Bridge R.: A test of Chargaff's second rule, *Biochemical and Biophysical Research Communications*, 340(1), 2006, 90-94.
15. Murphy W., Pringle T., Crider T., Springer M., Miller W.: Using genomic data to unravel the root of the placental mammal phylogeny, *Genome Research*, 17(4), 2007, 413-421.
16. Pearson H.: Genetics: What is a gene?, *Nature*, 441(7092), 2006, 398-401.
17. Prasad A., Allard M.: Confirming the phylogeny of mammals by use of large comparative sequence data sets, *Molecular Biology and Evolution*, 25(9), 2008, 1795-1808.
18. Seitz H.: *Analytics of protein-DNA interactions, Advances in Biochemical Engineering Biotechnology*, Berlin, Springer-Verlag, Heidelberg, 2007.
19. Schroeder M.: *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*, New York, W. H. Freeman, 1991.

20. Shepard R.N.: The analysis of proximities: Multidimensional scaling with an unknown distance function, *Psychometrika*, 27(I and II) (1962) 219-246 and 219-246.
21. Sims G., Jun S.R., Wu G., Kim S.H.: Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions, *Proc. of the National Academy of Sciences of the United States of America*, 106(8), 2009, 2677-2682.
22. Tuimala J.: *A primer to phylogenetic analysis using the PHYLIP package*, CSC - Scientific Computing Ltd., 2006.
23. Yu Z.-G., Anh V., Lau K.-S., Chu K.-H.: The genomic tree of living organisms based on a fractal model, *Physics Letters A*, 317(1), 1998, 293-302.
24. Yu Z.-G., Anh V., Lau K.-S.: Multifractal and correlation analyses of protein sequences from complete genomes, *Physical Review E*, 68(2), 2003, 021913.
25. Yu Z.-G., Anh V., Lau K.-S.: Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlational analysis, *Journal of Theoretical Biology*, 226(3), 2004, 341-348.
26. Zhang C.-T., Zhang R., Ou H.-Y.: The Z curve database: A graphic representation of genome sequences, *Bioinformatics*, 19(5), 2003, 593-599.
27. Zhao H., Bourque G.: Recovering genome rearrangements in the mammalian phylogeny, *Genome Research*, 19(5), 2009, 934-942.

José António Tenreiro Machado, Professor: Institute of Engineering, Polytechnic of Porto, Dept. of Electrical Engineering, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal, the author gave a presentation of this paper during one of the conference sessions (*jtm@isep.ipp.pt*).

Alexandra Maria Soares Ferreira Galhano, Professor: Institute of Engineering, Polytechnic of Porto, Dept. of Electrical Engineering, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal (*amf@isep.ipp.pt*).