Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2012, Article ID 132625, 12 pages
doi:10.1155/2012/132625

Research Article

Shannon Entropy Analysis of the Genome Code

J. A. Tenreiro Machado

Department of Electrical Engineering, Institute of Engineering of Polytechnic Porto, Rua Dr. António Bernardino de Almeida 431, 4200-072 Porto, Portugal

Correspondence should be addressed to J. A. Tenreiro Machado, jtm@isep.ipp.pt

Received 24 February 2012; Revised 8 April 2012; Accepted 10 April 2012

Academic Editor: Gerhard-Wilhelm Weber

This paper studies the chromosome information of twenty five species, namely, mammals, fishes, birds, insects, nematodes, fungus, and one plant. A quantifying scheme inspired in the state space representation of dynamical systems is formulated. Based on this algorithm, the information of each chromosome is converted into a bidimensional distribution. The plots are then analyzed and characterized by means of Shannon entropy. The large volume of information is integrated by averaging the lengths and entropy quantities of each species. The results can be easily visualized revealing quantitative global genomic information.

1. Introduction

Genome sequencing produced a huge volume of information that is now available for computational processing. Deschavanne et al. [1] explored DNA structures of genomes by means of a tool derived from the chaos dynamics. Murphy et al. [2] studied the genome sequences of four species to infer early events in placental mammal phylogeny. Ebersberger et al. [3] developed a phylogenetic analysis of several DNA sequence alignments from human, chimpanzee, gorilla, orangutan, and rhesus. Prasad et al. [4] analyzed a genomic sequence, which we generated from 41 mammals and 3 other vertebrates. In [5], Bolshoy reported a novel compositional complexity-based method for sequence analysis. The study shows that the method indicated periodicities and related features in several sets of DNA sequences. In [6], Liu et al. analyzed several aspects of the information content of the *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, and *Escherichia coli* genomes. In [7], Sims et al. used an alignment-free method in which l-mer frequency profiles of whole genomes are used for comparison. Macropol et al. [8] proposed an algorithm based on repeated random walks (RRWs) and apply the technique on a functional network of yeast genes identifying statistically significant clusters of

proteins. Kozobay-Avraham et al. [9] performed a genome analysis of DNA curvature distributions in coding and noncoding regions of prokaryotic genomes to evaluate the assistance of mathematical and statistical procedures. Two methods were applied producing similar clustering reflecting genomic attributes and environmental conditions of the species' habitat. Çarkacioğlu et al. [10] proposed the bi-k-bi clustering for finding association rules of gene pairs that can easily operate on large scale and multiple heterogeneous data sets. Kaplunovsky et al. [11] investigated correlations between certain properties of exons in a gene and genomic trees obtained with different approaches of clustering based on exonic parameters. They concluded that the best approach was based on distances among four principal components obtained by factor analysis, and followed by application of clustering algorithms. Sualp and Can [12] computed several graph theoretic measures on a protein-protein interaction network of a target organism as indicators of network context. Machado et al. [13] studied the human DNA from the perspective of system dynamics, associating entropy and the Fourier transform.

Based on the genomic data, this paper studies the deoxyribonucleic acid (DNA) code of twenty five species. Having in mind, the tools adopted in system and chaos analysis a state space representation and entropy measure are adopted. The state space plots reveal complex evolutions, resembling those revealed by chaotic systems and suggesting that the DNA information can be tackled by numerical tools. Given the large number of chromosomes and species involved in the study, the information is synthesized by means of the arithmetic averages of the entropy and the chromosome length. This strategy allows a simple quantitative visualization of the global genomic information of each species. Bearing these ideas in mind this paper is organized as follows. Section 2 presents the DNA code mapping concepts and the Shannon entropy characterization of the resulting numerical data. Section 3 analyzes the DNA entropy content of 489 chromosomes corresponding to twenty five species, including several mammals, fishes, birds, insects, nematodes, fungus, and one plant. Finally, Section 4 outlines the main conclusions.

2. Mapping the DNA Code and Quantification by Means of Entropy

The DNA helix encodes information by means of four distinct nitrogenous bases {thymine, cytosine, adenine, guanine} usually denoted by the symbols {T, C, A, G}. Besides the four symbols, the chromosome data files include a fifth symbol {N} which is believed to have no practical meaning for the DNA decoding. Each base connects with only one type of base on the other side forming the base pairing A-T and C-G.

The problem of DNA decoding is addressed in this paper using an algorithm inspired in system dynamical analysis using state space representation. This method was formulated by Roy et al. [14] and later addressed in conjunction with fractal dimension by Machado [15]. In the present paper, the scheme is improved by connecting the state plane with the entropy measure. The proposed strategy consists of implementing the translation scheme: (i) the A-T and C-G pairs are represented in the horizontal and vertical Cartesian axes, respectively, and (ii) each base along the DNA strand is converted to a one-step increment $\delta > 0$, being $+\delta$ ($-\delta$) for the first (second) base in each bonding pair. In the case of symbol {N}, no action is taken. Therefore, the DNA information, corresponding to the succession of bases, is converted into a trajectory representative of the dynamical evolution. Furthermore, the translation preserves the based pairing logic and does not introduce any preconception biasing the DNA information. In [15], it was adopted the box counting method for characterizing the fractal

image in the state plane. However, the box counting is an approximate method that requires large images in order to have a reasonable precision and does not quantify the case of successive trajectories passing through the same points. Having this fact in mind, in this paper, it is proposed an alternative method that takes into account the number of trajectories passing through a given point in the state plane. First, as in the case of using images, the minimum and maximum values along each axis are calculated and the trajectories are rescaled in order to fit a matrix $\mathbf{M}$ of size $n \times n$. Second, the points in the trajectories are quantified and counted for each cell in matrix $\mathbf{M}$. Third, the matrix $\mathbf{M}$ is converted to a bidimensional histogram by dividing each cell counting (that represents the number of trajectory points that fit inside the cell boundaries) by the total number of trajectory points.

The characterization of bidimensional histograms can be accomplished by several indices. In the paper, it is adopted the Shannon entropy [16–21]. Statistical indices based on moments can be used but that option requires a high number of measures. In fact, describing the frequency distribution of 25 species using the mean, variance, skewness, and kurtosis goes in the opposite direction of designing an assertive characterization and visualizing methodology. Furthermore, the histograms reveal irregular shapes, which preclude alleviating the total number and considering only a limited set of indices.

The concept of entropy was developed by Ludwig Boltzmann when analyzing the statistical behavior of system's microscopic components. In information theory, entropy was devised by Claude Shannon to study the amount of information in a transmitted message. The Shannon entropy H, satisfying the Shannon-Khinchin axioms, is defined as

$$H(X) = -\sum_{x \in X} p(x) \ln\left[p(x)\right], \tag{2.1}$$

where $p(x)$ is the probability that event $x \in X$ occurs.

For bidimensional probability distributions, the expression becomes

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \ln\left[p(x,y)\right], \tag{2.2}$$

where $p(x,y)$ is the joint probability distribution function of (X,Y).

The entropy index H is applied to 25 species having the main characteristics depicted in Table 1 and totalizing 489 chromosomes.

3. DNA Entropy and Chromosome Length

The code in each of the 489 chromosomes is converted to a state plane portrait, and the bidimensional histogram is described in the light of the entropy measure. Several experiments varying the number of cells of the $n \times n$ matrix $\mathbf{M}$ demonstrated that there are only minor numerical differences once large values are adopted, and it was found that $n = 100$ is a good compromise between precision and computational requirements.

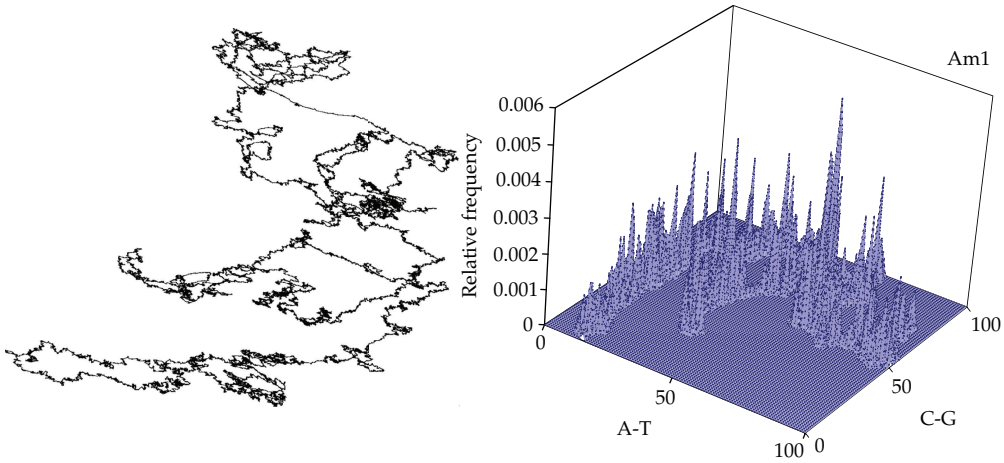
Figure 1 shows, for example, the two-dimensional state plane plots and the corresponding bidimensional distribution of relative frequency of the chromosomes Am1, Hu1, Tg1, and Zf1. The horizontal and vertical axes are not represented since they have no useful contribution for the calculations.

Table 1: Species and chromosomes.

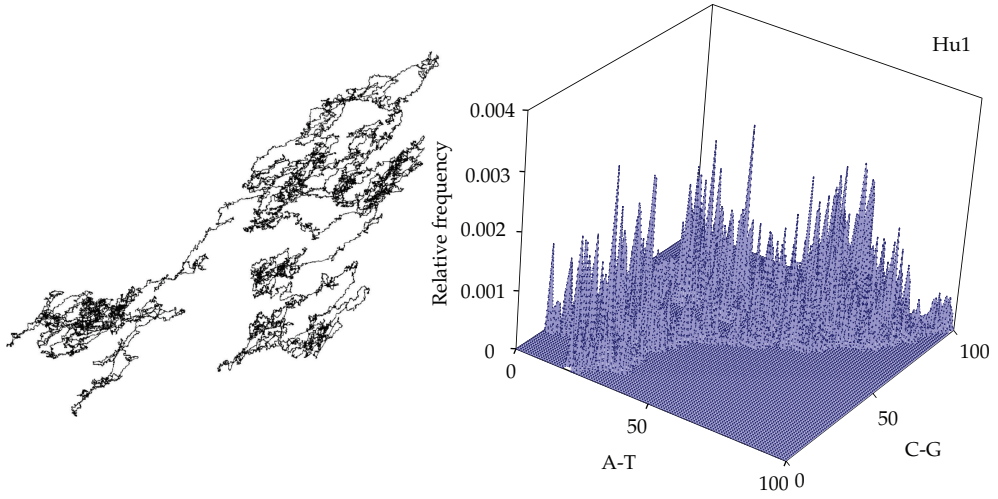
Species	Tag	Group	Number of chromosomes	Arithmetic average of logarithm of chromosome length Av[ln(L)]	Arithmetic average of entropy Av(H)
Mosquito (*Anopheles gambiae*)	Ag	Insect	5	17.59	7.57
Honeybee (*Apis mellifera*)	Am	Insect	16	16.24	6.92
Arabidopsis thaliana	At	Plant	5	16.97	6.81
Caenorhabditis briggsae	Cb	Nematode	6	16.64	7.07
Caenorhabditis elegans	Ce	Nematode	6	16.64	7.07
Chimpanzee	Ch	Mammal	25	18.54	7.14
Dog	Dg	Mammal	39	17.90	6.94
Drosophila simulans	Ds	Insect	6	16.29	7.36
Drosophila yakuba	Dy	Insect	10	14.57	6.93
Horse	Eq	Mammal	32	18.02	7.02
Chicken	Ck	Bird	31	15.64	6.27
Human	Hu	Mammal	24	18.58	7.13
Medaka	Me	Fish	24	17.23	7.25
Mouse	Mm	Mammal	21	18.57	7.16
Opossum	Op	Mammal	9	19.63	7.47
Orangutan	Or	Mammal	24	18.57	7.12
Cow	Ox	Mammal	30	18.24	7.08
Pig	Po	Mammal	19	18.52	6.85
Rhesus	Rm	Mammal	21	18.69	6.99
Rat	Rm	Mammal	21	18.59	7.20
Yeast (*Saccharomyces cerevisiae*)	Sc	Fungus	16	13.41	6.91
Stickleback	St	Fish	21	16.75	7.10
Zebra Finch	Tg	Bird	32	16.18	6.31
Tetraodon	Tn	Fish	21	16.19	7.14
Zebrafish	Zf	Fish	25	17.79	6.83

The charts of the 489 chromosomes were analyzed, and it was concluded that (i) the plots vary considerably and are a signature of each case, (ii) there were significant areas of the state plane that were not visited by the trajectories, and (iii) there were parts of the charts constituted by lines or by part of lines along the +45 or −45 degree direction.

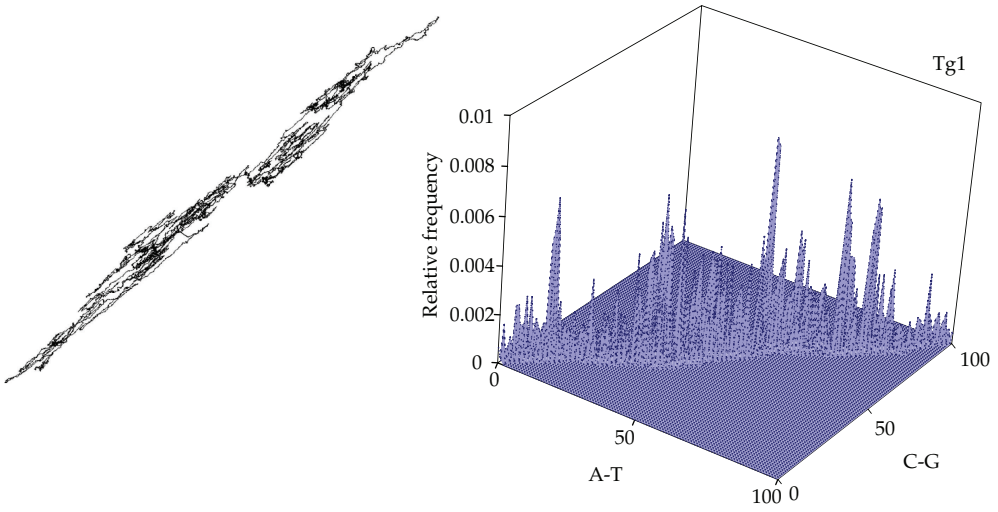
For each chromosome, the Shannon entropy was calculated. For example, in the bidimensional histograms of Figure 1 were obtained the values $H_{\mathrm{Am1}} = 7.092$, $H_{\mathrm{Hu1}} = 7.242$, $H_{\mathrm{Tg1}} = 6.240$, and $H_{\mathrm{Zf1}} = 6.676$.



(a)

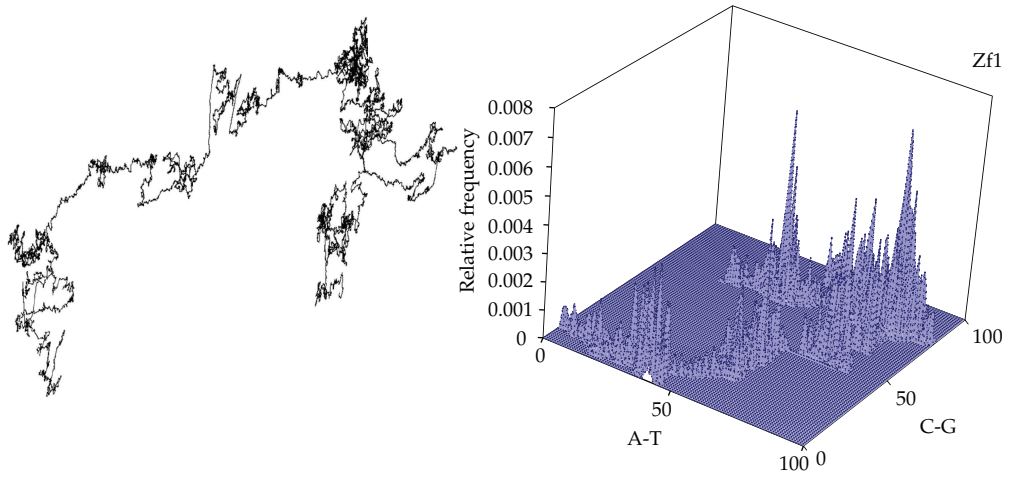


(b)



(c)

Figure 1: Continued.



(d)

Figure 1: State plane portraits and relative frequency distribution of the chromosomes: (a) Am1, (b) Hu1, (c) Tg1, (d) Zf1.

The quality of the entropy index was verified by two sets of experiments, namely, by comparing it with two alternative measures, and by assessing three artificial test files. In the first set of experiments, the fractal dimension (FD) of the two-dimensional state portraits and the Mutual information (I) of the bidimensional relative frequency distribution were calculated as alternative measures.

For estimating the fractal dimension, the box counting method was adopted [22–24]. For a set S in a n-dimensional and any $\varepsilon > 0$, if there is a number FD so that $N_\varepsilon(S) \sim 1/\varepsilon^{\mathrm{FD}}$ as $\varepsilon \to 0$, where $N_\varepsilon(S)$ is the minimum number of n-dimensional cubes of side-length ε needed to cover S, we say that the box counting dimension of S is FD. This reasoning leads to the expression:

$$\mathrm{FD} = -\lim_{\varepsilon \to 0} \frac{\ln[N_\varepsilon(S)]}{\ln(\varepsilon)}, \tag{3.1}$$

which can be implemented with image processing algorithms.

In our case, S consists of the state plane monochrome images and small values of ε are reached by accessing images at the pixel level.

The mutual information (I) of two random variables measures the dependence between two random variables and is defined as

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \ln \frac{p(x,y)}{p(x)p(\mathrm{y})}, \tag{3.2}$$

where $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y, respectively.

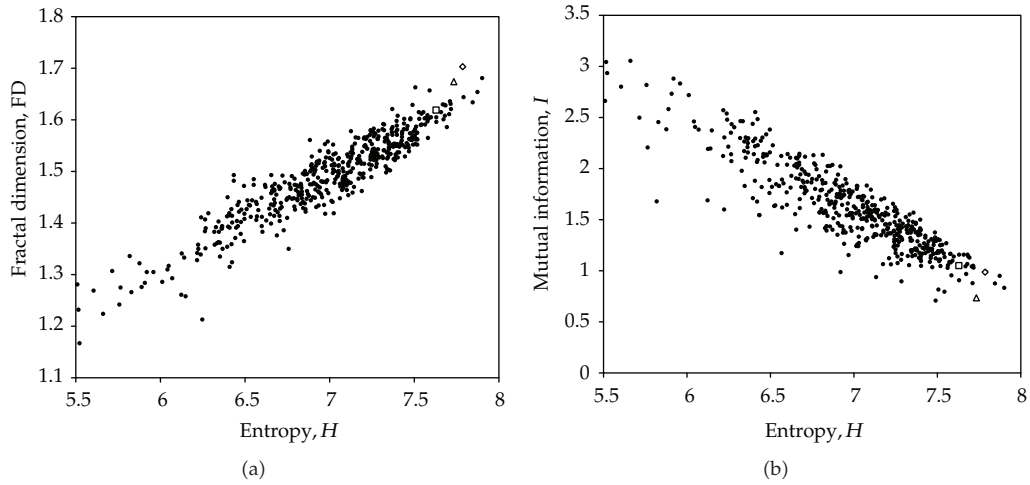


Figure 2: Entropy H versus (a) fractal dimension FD, (b) mutual information I. The black circles represent the 489 chromosomes. The white markers, namely, the square, triangle, and diamond, represent the test sequences of 1, 2 and 3 symbols.

The mutual information can be expressed as

$$I(X,Y) = H(X) + H(Y) - H(X,Y), \tag{3.3}$$

where $H(X)$ and $H(Y)$ are the marginal entropies.

In the second set of experiments three files with random permutations of 1, 2, and 3 symbol sequences of {T, C, A, G} were generated and treated as if they were chromosome files. For these files, the probabilities are identical (i.e., 1/4, 1/16 and 1/64, for the 1, 2, and 3 symbol sequences) and the number of generating iterations adjusted so that they were 1 megabyte length.

Figure 2 shows the entropy H of the state plane histogram versus the fractal dimension FD and the mutual information I. The black circles represent the 489 chromosomes, while the white markers at the right corners, namely, the square, the triangle, and the diamond, represent the test sequences of 1, 2, and 3 symbols. We verify that there is a strong correlation in both cases, and, therefore, results are expected to be qualitatively of the same type. In the case of FD, this is due to the fact that the relative frequency distribution concentrates into a few spots, making the information along the z-axis less significant than the one represented by the x- and y-axes. Nevertheless, also due to that same reason, H is slightly superior to FD. Since the relative frequency of the four symbols is approximately identical, in the case of I the marginal entropies are almost constant and expression (3.1) leads to a linear relationship with H. In what concerns the three test files, we observe the white markers are located at the right limits of the set of points, and, consequently, the proposed scheme is capable of distinguishing between the natural and the artificial data files.

Figure 3 shows the relationship between the entropy H and the length L of the 489 chromosomes. Analyzing individually each of the species we observe some grouping that reflects the qualitative analysis held initially for each separate plot. For each species, an individual map can be plotted, showing the relative similarities of the chromosomes. For example, Figure 4 represents the *locus* of H versus L for the 24 and the 16 chromosomes of

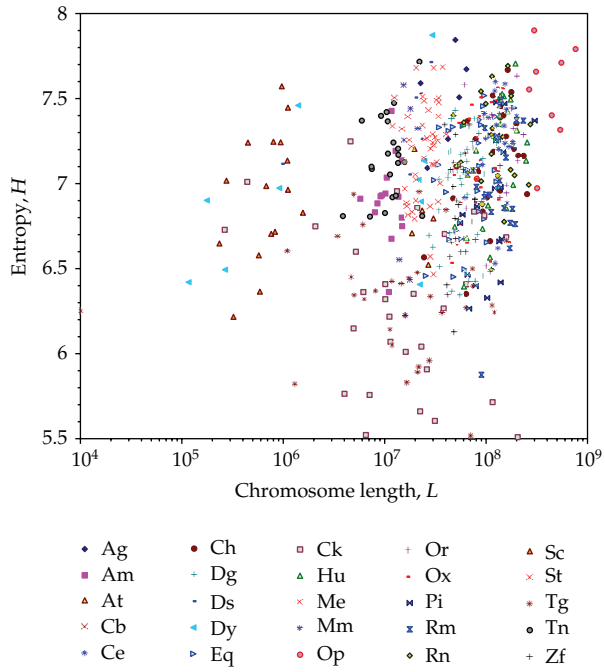


Figure 3: Entropy H versus chromosome length L for the 25 species.

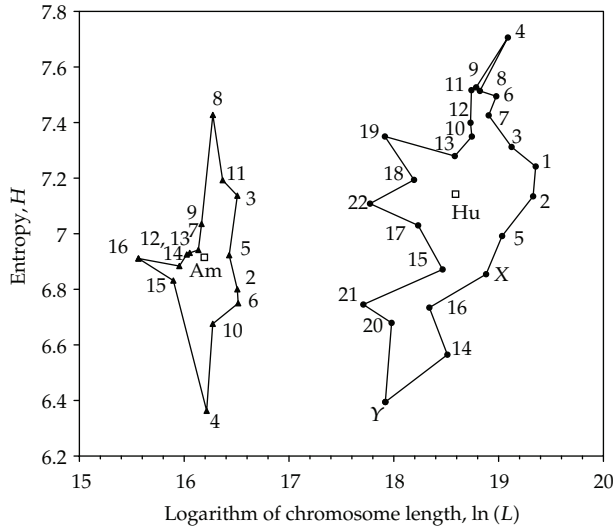


Figure 4: Entropy H versus logarithm of chromosome length $\ln(L)$ the 24 and the 16 chromosomes of Hu and Am, respectively. The white markers represent the arithmetic average of the horizontal and vertical coordinates for each set of chromosomes.

Hu and Am, respectively. The white markers represent the arithmetic average of the horizontal and vertical coordinates for each species and can be interpreted as the "center" of each set of chromosomes. For the Hu, we observe that the chromosomes 4 and Y are in opposite parts of the set, while, for the Am, chromosomes 4 and 8 are the most distant ones.

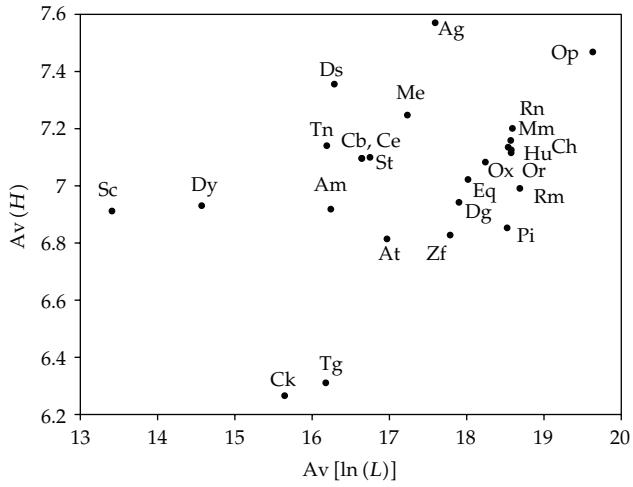


Figure 5: Arithmetic averages of entropy versus logarithm of chromosome length, Av(H) versus Av[ln(L)], for the 25 species.

Figure 3 includes a considerable number of points, and, therefore, some sort of integration action is necessary. In this perspective, for each species, the arithmetic averages of the entropy and the logarithm of chromosomes lengths (i.e., Av(H) versus Av[ln(L)]) are applied. The plot depicted in Figure 5 reveals the emergence of patterns that are in accordance with phylogenetics. The corresponding numerical values are depicted in the two right columns of Table 1.

At the left are located the less complex species and at right are plotted the mammals. Within the cluster of mammals, the primates {Ho, Ch, Or} form a subcluster. Among the mammals, it is interesting to notice Mm close to the primates and the extreme position of the marsupial Op, relatively distant from the placental mammals. In what concerns the remaining points, we verify Cb to be almost indistinguishable from Ce. In a middle position, we have the clusters of birds {Ck, Tg} and fishes {Tn, St, Me, Zf}. It is interesting to see that the plant At is located between the insect Am and the fish Zf. Finally, at the extreme left, we have Sc.

Since the mammals have a relative close position in a narrow region of the map, it is important to analyze the zoom represented in Figure 6 where it is clear the close position not only of the primates {Ho, Ch, Or} but also of Mm and Rn.

4. Conclusions

Chromosomes have a code based on a four-symbol alphabet, and the information can be analyzed with tools adopted in dynamical systems. In this paper, a translation scheme for converting the DNA sequence into a state plane trajectory was adopted. The application to the 489 data files of 25 species revealed bidimensional histograms representative of each chromosome. The results were processed by means of Shannon entropy, and, in order to obtain a simple visualization, the values were averaged for each species. The map of entropy versus chromosome length revealed the emergence of comprehensive patterns of the species relative characteristics. It was verified that the mammals form a cluster located in a narrow area of

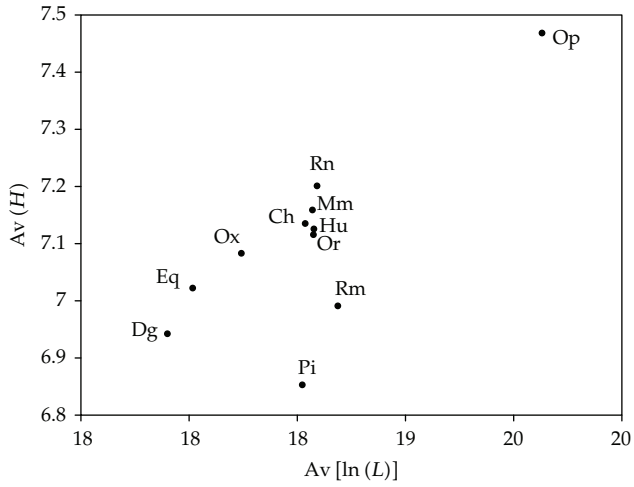


Figure 6: Averages of entropy versus logarithm of chromosome length, Av(H) versus Av[ln(L)], for the 11 mammals.

the map and that the mouse and rat are relatively close to the primates, while the marsupial is far from the rest of the placental species.

Acknowledgments

This work is supported by FEDER Funds through the "Programa Operacional Factores de Competitividade—COMPETE" program and by National Funds through FCT "Fundação para a Ciência e a Tecnologia" under the Project FCOMP-01-0124-FEDER-PEst-OE/EEI/UI0760/2011. The author thanks the following organizations for allowing access to genome data: Human, Genome Reference Consortium, http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/; Common Chimpanzee, Chimpanzee Genome Sequencing Consortium; Orangutan, Genome Sequencing Center at WUSTL, http://genome.wustl.edu/genome.cgiGENOME=Pongo%20abelii; Rhesus, Macaque Genome Sequencing Consortium, http://www.hgsc.bcm.tmc.edu/projects/rmacaque/; Pig, The Swine Genome Sequencing Consortium, http://piggenome.org/; Opossum, The Broad Institute, http://www.broad.mit.edu/mammals/opossum/; Chicken, International Chicken Genome Sequencing Consortium Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004 Dec 9; 432(7018): 695-716. PMID: 15592404; Zebra Finch, Genome Sequencing Center at Washington University St. Louis School of Medicine; Zebrafish, The Wellcome Trust Sanger Institute, http://www.sanger.ac.uk/Projects/D_rerio/; Tetraodon, Genoscope, http://www.genoscope.cns.fr/; Honeybee, The Baylor College of Medicine Human Genome Sequencing Center, http://www.hgsc.bcm.tmc.edu/projects/honeybee/; Gambiae Mosquito, The International Anopheles Genome Project; Elegans nematode, Wormbase, http://www.wormbase.org/; Briggsae nematode, Genome Sequencing Center at Washington University in St. Louis School of Medicine; Yeast, Saccharomyces Genome Database, http://www.yeastgenome.org/; Dog, http://www.broad.mit.edu/mammals/dog/; *Drosophila simulans*, http://genome.wustl.edu/genomes/view/drosophila_simulans_white_501; *Drosophila yakuba*, http://genome.wustl.edu/genomes/view/

drosophila_yakuba; Horse, http://www.broad.mit.edu/mammals/horse/; Medaka, http://dolphin.lab.nig.ac.jp/medaka/; Mouse Genome Sequencing Consortium, http://www.hgsc.bcm.tmc.edu/projects/mouse/; Ox, The Baylor College of Medicine Human Genome Sequencing Center, http://www.hgsc.bcm.tmc.edu/projects/bovine/; Rat, The Baylor College of Medicine Human Genome Sequencing Center, http://www.hgsc.bcm.tmc.edu/projects/rat/; Stickleback, http://www.broadinstitute.org/scientific-community/science/projects/mammals-models/vertebratesinvertebrates/stickleback/stickleba; The Wellcome Trust Sanger Institute, http://www.sanger.ac.uk/Projects/D_rerio/; The Arabidopsis Information Resource, http://www.arabidopsis.org/.

References

[1] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertit, "Genomic signature: characterization and classification of species assessed by chaos game representation of sequences," *Molecular Biology and Evolution*, vol. 16, no. 10, pp. 1391–1399, 1999.

[2] W. J. Murphy, T. H. Pringle, T. A. Crider, M. S. Springer, and W. Miller, "Using genomic data to unravel the root of the placental mammal phylogeny," *Genome Research*, vol. 17, no. 4, pp. 413–421, 2007.

[3] I. Ebersberger, P. Galgoczy, S. Taudien, S. Taenzer, M. Platzer, and A. Von Haeseler, "Mapping human genetic ancestry," *Molecular Biology and Evolution*, vol. 24, no. 10, pp. 2266–2276, 2007.

[4] A. B. Prasad, M. W. Allard, and E. D. Green, "Confirming the phylogeny of mammals by use of large comparative sequence data sets," *Molecular Biology and Evolution*, vol. 25, no. 9, pp. 1795–1808, 2008.

[5] A. Bolshoy, "Revisiting the relationship between compositional sequence complexity and periodicity," *Computational Biology and Chemistry*, vol. 32, no. 1, pp. 17–28, 2008.

[6] Z. Liu, S. S. Venkatesh, and C. C. Maley, "Sequence space coverage, entropy of genomes and the potential to detect non-human DNA in human samples," *BMC Genomics*, vol. 9, p. 509, 2008.

[7] G. E. Sims, S. R. Jun, G. A. Wu, and S. H. Kim, "Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 8, pp. 2677–2682, 2009.

[8] K. Macropol, T. Can, and A. K. Singh, "RRW: repeated random walks on genome-scale protein networks for local cluster discovery," *BMC Bioinformatics*, vol. 10, p. 283, 2009.

[9] L. Kozobay-Avraham, S. Hosid, Z. Volkovich, and A. Bolshoy, "Prokaryote clustering based on DNA curvature distributions," *Discrete Applied Mathematics*, vol. 157, no. 10, pp. 2378–2387, 2009.

[10] L. Çarkacioğlu, R. Atalay, Ö. Konu, V. Atalay, and T. Can, "Bi-k-bi clustering: mining large scale gene expression data using two-level biclustering," *International Journal of Data Mining and Bioinformatics*, vol. 4, no. 6, pp. 701–721, 2010.

[11] A. Kaplunovsky, A. Ivashchenko, and A. Bolshoy, "Statistical analysis of exon lengths in various eukaryotes," *Open Access Bioinformatics*, vol. 3, pp. 1–15, 2011.

[12] M. Sualp and T. Can, "Using network context as a filter for miRNA target prediction," *BioSystems*, vol. 105, no. 3, pp. 201–209, 2011.

[13] J. A. T. Machado, A. C. Costa, and M. D. Quelhas, "Entropy analysis of the DNA code dynamics in human chromosomes," *Computers & Mathematics with Applications*, vol. 62, no. 3, pp. 1612–1617, 2011.

[14] 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.

[15] J. T. Machado, "Accessing complexity from genome information," *Communications in Nonlinear Science and Numerical Simulations*, vol. 17, no. 6, pp. 2237–2243, 2012.

[16] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[17] E. T. Jaynes, "Information theory and statistical mechanics," vol. 106, no. 4, pp. 620–630, 1957.

[18] A. I. Khinchin, *Mathematical Foundations of Information Theory*, Dover Publications, New York, NY, USA, 1957.

[19] C. Beck, "Generalised information and entropy measures in physics," *Contemporary Physics*, vol. 50, no. 4, pp. 495–510, 2009.

[20] R. M. Gray, *Entropy and Information Theory*, Springer-Verlag, New York, NY, USA, 2009.

[21] J. T. Machado, "Entropy analysis of integer and fractional dynamical systems," *Nonlinear Dynamics*, vol. 62, no. 1-2, pp. 371–378, 2010.

[22] M. V. Berry, "Diffractals," *Journal of Physics A*, vol. 12, no. 6, pp. 781–797, 1979.
[23] M. L. Lapidus and J. Fleckinger-Pellé, "Tambour fractal: vers une résolution de la conjecture de Weyl-Berry pour les valeurs propres du laplacien," *Comptes Rendus des Séances de l'Académie des Sciences*, vol. 306, no. 4, pp. 171–175, 1988.
[24] M. Schroeder, *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*, W. H. Freeman and Company, New York, NY, USA, 1991.