Accessing complexity from genome information

J.A.TenreiroMachado

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

ABSTRACT

This paper studies the information content of the chromosomes of 24 species. In a first phase, a scheme inspired in dynamical system state space representation is developed. For each chromosome the state space dynamical evolution is shed into a two dimensional chart. The plots are then analyzed and characterized in the perspective of fractal dimen- sion. This information is integrated in two measures of the species' complexity addressing its average and variability. The results are in close accordance with phylogenetics pointing quantitative aspects of the species' genomic complexity.

Keywords:
DNA, Chromosome Fractals Information Phylogenetics

1. Introduction

The genome sequencing produced considerable information that is presently available for analytical and computational processing [1–13]. This paper addresses the code information embedded in the deoxyribonucleic acid (DNA) of 24 species. During the last years several researcher have tackled the issue of genome complexity [14–16] but the fact is that many ques- tions remain open. Having in mind the tools adopted in system modeling and chaos analysis, in this paper several tools, namely state space graphical representation and fractal dimension are adopted. In fact, the state space charts reveal complex evolutions, having similarities with those depicted by chaotic systems, suggesting that the DNA information can by tackled by standard analytical and numerical methods. Given the large number of chromosomes, two synthesizing and comparison indices based on the average and variability of the fractal dimension and chromosome lengths are developed. These mea- sures lead to a clear map of species not only in accordance with known phylogenetics, but also with quantitative assessment of the complexity.

Having these ideas in mind, this paper is organized as follows. Section 2 presents the DNA sequence decoding concepts, the mathematical tools, and formulates the indices that reflect the complexity content and variability of each species. Section 3 analyzes the DNA information content of 489 chromosomes corresponding to a set of 24 species. Finally, Section 4 outlines the main conclusions.

2. DNA and information analysis

In the DNA double helix there are four distinct nitrogenous bases, namely thymine, cytosine, adenine and guanine, usually denoted by the symbols {T, C, A, G}. Each type of base on one strand connects with only one type of base on the other strand, forming the base pairing A–C and T–G. Besides the four symbols {T, C, A, G}, the available chromosome data includes a fifth symbol ‘‘N’’ which is believed to have no practical meaning for the DNA decoding.

The DNA information decoding constitutes a formidable challenge and this paper addresses this issue inspired in (i) dynamical systems modeling using state space representation, (ii) chaos analysis using fractal dimension concepts, and
(iii) information measures.

Dynamical systems are assertively described using the so-called state space modeling. For that purpose it is necessary to start by defining the type and number of state variables. They represent the systems’ fundamental ingredients and its dynamics can be evaluated based in time evolution of the state variables. Often two dimensional models that lead to theso-called state plane are adopted, allowing direct graphical representations.

Bearing these ideas in mind it was decided to have a two-dimensional state space representation of the DNA information based on a simple translation scheme. First the A–C and T–G pairs are represented in the horizontal and vertical Cartesian axes, respectively. Second, each base along the DNA strand is converted to a one-step increment d, being d > 0 (d < 0) for the first (second) base in each bonding pair. In the case of symbol ‘‘N’’ no action is taken. By other words, representing by *x* and *y* the horizontal and vertical coordinates, for each symbol read along the sequence it is adopted one iteration step of the type: hA, T**i**: *x* ? h*x* + d, *x* – d**i** or hC, G**i**: *y* ? h*y* + d, *y* – d**i**. For example, with d = 1 when starting from (0, 0), the code {ACA- CACACTTGTGTGG} translates to the Cartesian coordinates (*x*, *y*) = (0, 0), (1, 0), (1, 1), (2, 1), (2, 2), (2, 3), (3, 3), (4, 3), (4, 4),
(3, 4), (2, 4), (2, 3), (1, 3), (1, 2), (0, 2), (0, 1), (0, 0) in the state space. Therefore, the succession of bases is converted to a chartrepresentative of the dynamical evolution that can be analyzed with mathematical tools usual in system theory. Further-more, the translation scheme preserves the based pairing logic and does not introduce any preconception biasing theDNA information.

It should be noted that according with the second Chargaff’s rule the number of symbols A and T, and G and C are approx- imately identical, not only for each of the two DNA strands, but also for long sequences [17–19]. Nevertheless, in the presentcase we are capturing the order of the symbols along the sequence and, therefore, considerable deviations from the 45° lineoccur. Computation of the complexity for DNA representations is interesting and we can also mention the *Z*-curve [20].

The second phase consists of extracting information from the two dimensional state space charts. Since the results, to be analyzed in the next section, have close resemblances to those of chaotic systems it was chosen the box-counting method forcharacterizing the plots [21–23].

The box-counting dimension of a set *S* in a *n*-dimensional space is defined as follows: for any e > 0, let *N*e(*S*) be the min- imum number of *n*-dimensional cubes of side-length e needed to cover *S*. If there is a number *d* so that
e^d
*N*eð*S*Þ rv 1 as e ? 0 we
say that the box-counting dimension of *S* is *d*. This reasoning leads to the expression:

$$d = -\lim_{\varepsilon\to 0} \frac{\ln\left[N_\varepsilon(S)\right]}{\ln(\varepsilon)} \tag{1}$$

which can be easily implemented with computational methods.

Table 1
Species, chromosome and main characteristics.

i	Species	Tag	Group	*Ni*
1	Mosquito (*Anopheles*	Ag	Insect	6
2	Honeybee (*Apis mellifera*)	Am	Insect	16
3	*Caenorhabditis briggsae*	Cb	Nematode	6
4	*Caenorhabditis elegans*	Ce	Nematode	6
5	Chimpanzee	Ch	Mammal	25
6	Dog	Dg	Mammal	39
7	*Drosophila simulans*	Ds	Insect	7
8	*Drosophila yakuba*	Dy	Insect	11

9	Horse	Eq	Mammal	32
10	Chicken	Ga	Bird	30
11	Human	Ho	Mammal	24
12	Medaka	Me	Fish	25
13	Mouse	Mm	Mammal	21
14	Opossum	Op	Mammal	9
15	Orangutan	Or	Mammal	24
16	Cow	Ox	Mammal	30
17	Pig	Po	Mammal	19
18	Rat	Rn	Mammal	21
19	Rhesus	Rm	Mammal	21
20	Yeast *(Saccharomyces*	Sc	Fungus	16
21	Stickleback	St	Fish	22
22	Zebra Finch	Tg	Bird	31
23	Tetraodon	Tn	Fish	21
24	Zebrafish	Zf	Fish	25

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

The third phase consists of integrating the fractal measures of the chromosome in order to establish an index representative of the complexity of each species. On one hand, we should note that a high/low fractal dimension represents a rich/ poor dynamical behavior, where rich/poor can be interpreted as complex/simple. On the other hand, we verify that species exhibit distinct number of chromosomes and different chromosomes' lengths that must reflect upon the total amount of information, but that those numbers vary significantly. Therefore, for the species *j* it is considered the complexity average *cj* defined as:

$$c_j = \left[\sum_{i=1}^{N_j} d_{ji} \ln(l_{ji})\right]^{1/\sum_{i=1}^{N_j} d_{ji}} \qquad (2)$$

where the index *i* represents the chromosome, *Nj* is the total number of chromosomes of species *j* and *dji* denotes the cor- responding fractal dimension.

This expression is inspired in the generalized average formulae. In this line of thought, it is relevant to measure not only the average value, but also the complexity variability between the set of chromosomes using the index *vj* defines as:

$$v_j = \sqrt{\left[\sum_{i=1}^{N_j} d_{ji} \left[\ln(l_{ji})\right]^2\right]^{1/\sum_{i=1}^{N_j} d_{ji}} - c_j^2} \qquad (3)$$

These analytical indices are applied to a set of 24 species having the main characteristics depicted in Table 1.

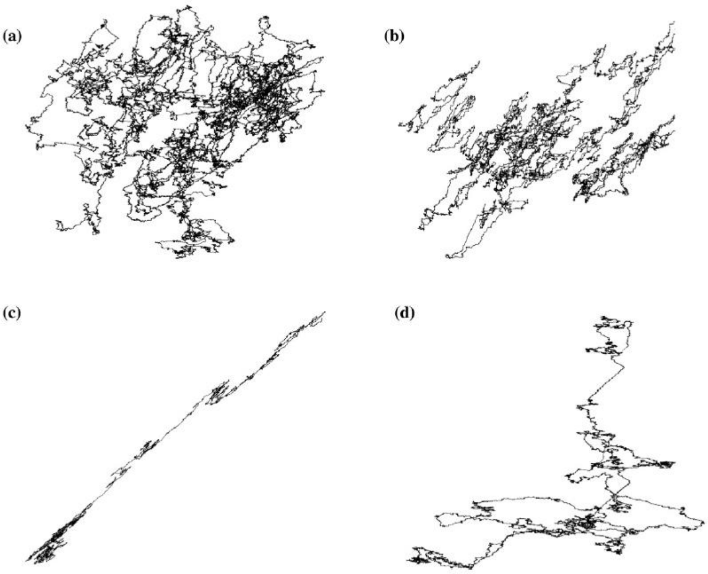

Fig. 1. Phase plane portraits of chromosomes: (a) Ag2L: $l_{1,1} = 49770995$, $d_{1,1} = 1.634$, (b) Eq1: $l_{9,1} = 189554878$, $d_{9,1} = 1.529$, (c) Ga1: $l_{10,1} = 205013902$, $d_{10,1} = 1.281$, (d) Sc1: $l_{20,1} = 234819$, $d_{20,1} = 1.417$.

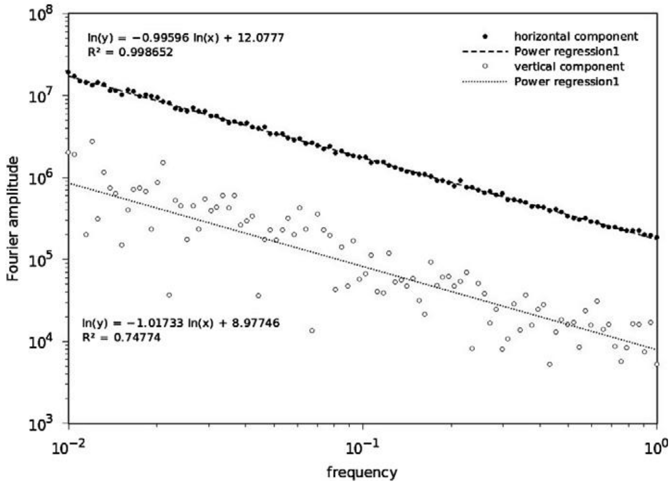

Fig. 2. Amplitude of the Fourier content of the phase plane plot for chromosome Eq1.

3. DNA information and species complexity

The 24 species totalize 489 chromosomes. Each of these chromosomes is analyzed for extracting the corresponding fractal dimension of the state plane portrait. Therefore, in a first step the chromosome information is read and the maximum and minimum limits of the state plane trajectory, along both axes, are evaluated. This preliminary evaluation allows the calcu- lation of a scale factor so that the final chart and the corresponding bitmap file have identical dimension regardless of the chromosome length. Therefore, we have the guarantee that the calculation of the fractal dimension is only a result of the DNA information content. Having calculated the limits and scale factor, the chromosome is read a second time and the state plane trajectory is plotted.

For example Fig. 1 shows the phase plane plots of the chromosomes Ag2L, Ga1, Eq1, and Sc1. The horizontal and vertical axes are not represented since they have no useful contribution for the calculations.

The plots vary considerably suggesting that they are sensitive to the code and their characteristics. For the examples of Fig. 1 we get the values Ag2L: $l_{1,1}$ = 49770995, $d_{1,1}$ = 1.634, Eq1: $l_{9,1}$ = 189554878, $d_{9,1}$ = 1.529, Ga1: $l_{10,1}$ = 205013902, $d_{10,1}$ = 1.281, and Sc1: $l_{20,1}$ = 234819, $d_{20,1}$ = 1.417. Moreover, it was observed some consistency of plots for the types of spe- cies. While this approach leads only to a qualitative analysis, it was verified that the application of the fractal dimension (1) was consistent with the observation and lead to a quantitative measure.

The charts have strong resemblances to those of random walks and Levy flights. Therefore, the Fourier transform was cal- culated for the *x* and *y* components of the image. For example, Fig. 2 depicts the amplitude of the Fourier content of the phase

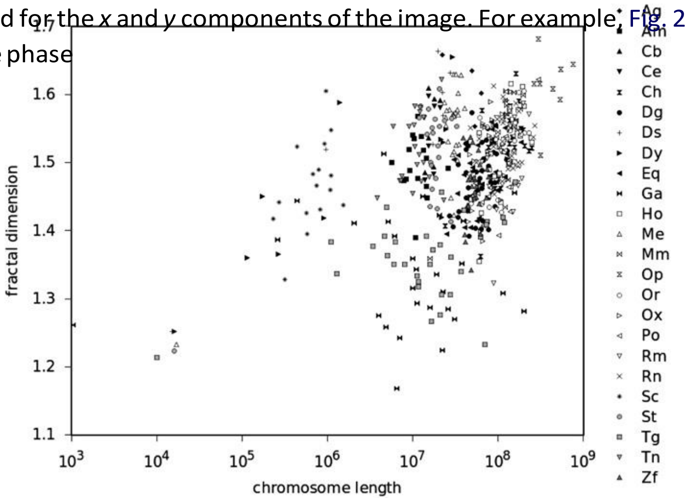

Fig. 3. Chromosome length versus fractal dimension of the state plane chart for the 24 species.

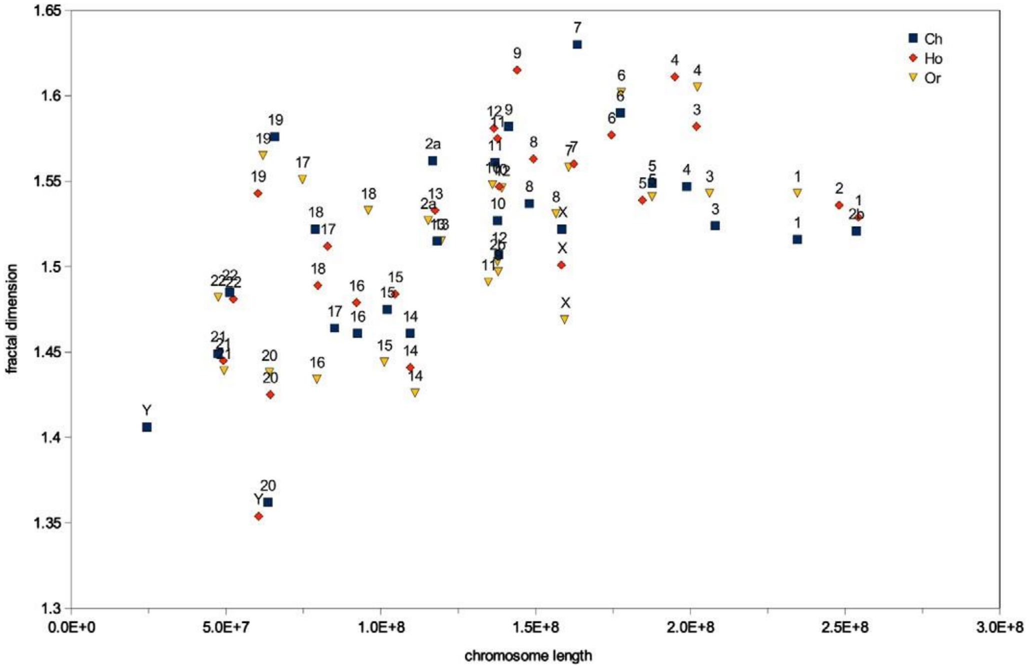

Fig. 4. Chromosome length versus fractal dimension of the state plane chart for the primates {Ch, Ho, Or}.

plane plot of chromosome Eq. (1) of Fig. 1b. The results can be easily approximated by a power law expression of the type *amplitude* rv *a*x^{b}; *a*; *b* 2 R, where x can loosely be denoted as ‘‘frequency’’ if we consider that each step d is a ‘‘time’’ increment. It was observed that *a* varied from chart to chart, depending on the chromosome and the horizontal/vertical compo- nent, but *b* remained almost invariant namely as *b* � −1.

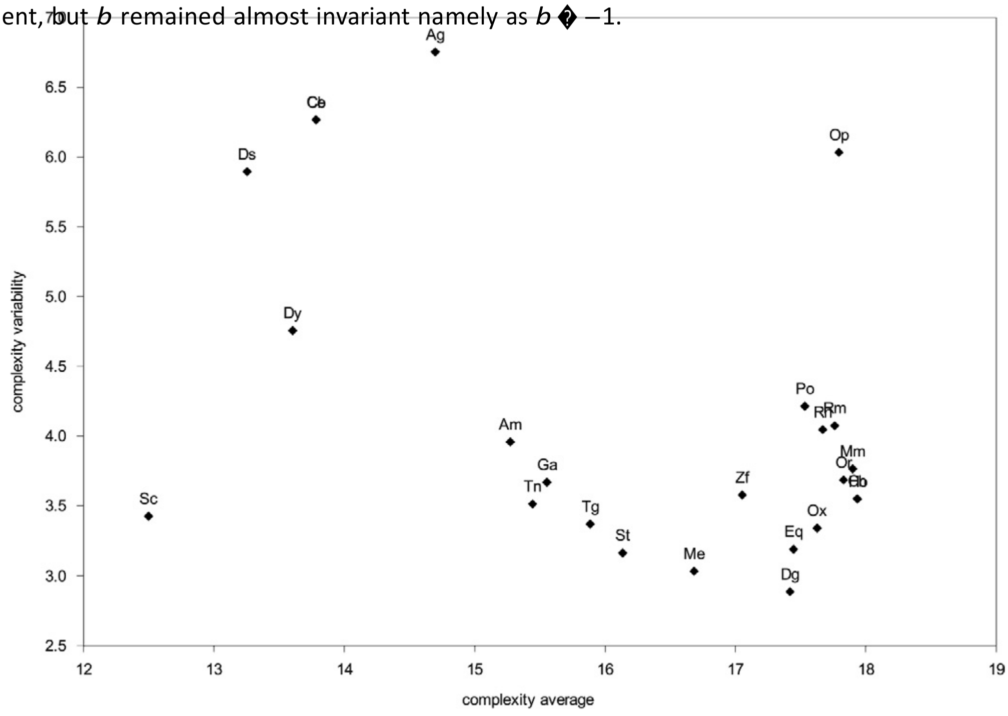

Fig. 5. Complexity average cj versus variability vj for the 24 species.

Table 2
Species' complexity.

i	Tag	*ci*	*vj*
1	Ag	14.698	6.751
2	Am	15.272	3.958
3	Cb	13.784	6.265
4	Ce	13.783	6.266
5	Ch	17.934	3.550
6	Dg	17.419	2.887
7	Ds	13.254	5.894
8	Dv	13.605	4.757
9	Eq	17.445	3.191
10	Ga	15.554	3.670
11	Ho	17.937	3.550
12	Me	16.683	3.034
13	Mm	17.900	3.765
14	Op	17.794	6.032
15	Or	17.829	3.686
16	Ox	17.626	3.340
17	Po	17.531	4.214
18	Rn	17.762	4.074
19	Rm	17.670	4.046
20	Sc	12.499	3.425
21	St	16.134	3.164
22	Tg	15.887	3.369
23	Tn	15.444	3.513
24	Zf	17.052	3.577

Fig. 3 shows the relationship between the fractal dimension and the length of the chromosomes. We observe the emergence of some grouping reflecting the qualitative analysis held initially for each separate plot. Parts of this map can beezoomed and the relationship between individual chromosomes can be visualized. For example Fig. 4 depicts the map forthe primates {Ch, Ho, Or}. Nevertheless, while these charts constitute a quantitative evaluation, the fact is that we have still a considerable amount of cases and the application of some sort of integration measure is advisable.

The application of indices (2) and (3) upon the 24 species produces the map of complexity average *cj* versus variability *vj*
depicted in Fig. 5 and to the list the values of Table 2.

We verify the emergence of patterns that are in accordance with phylogenetics, going from the less complex species Sc, at left, up to the most complex species Hu, at the right. The cluster of mammals is at right and, within it, the sub-cluster of primates {Ho, Ch, Or} with the Ch closer to Hu than the Or. In the rest of mammals it is interesting to note Mm close to the primates and the position of the marsupial Op relatively distant from the placental mammals in terms of complexity var- iability of the chromosomes. In what concerns the rest of the points we verify Cb to be almost indistinguishable from Ce that, together with the group of insects, reveal a low average but a high variability of complexity. In a middle position, with med- ium complexity but low variability (similar to the mammals) we have the clusters of birds {Ga, Tg} and fishes {Tn, St, Me, Zf}.

In conclusion, the proposed complexity measures lead to assertive quantitative classification of chromosomes and species.

4. Conclusions

Chromosomes have a code based on a four symbol alphabet. This information can be analyzed with tools usually adopted in dynamical system signal modeling. In this paper it was proposed a conversion scheme translating the DNA sequence to a phase plane chart. The application to the chromosomes of 24 species revealed patterns typical in chaotic systems. Bearing these facts in mind, the images were processed and the resulting values were embedded into two complexity measures based on the chromosome length and state space fractal dimension. The resulting map revealed the emergence of clear pat- terns capable of being interpreted and compared.

Acknowledgments

We thank 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 gen- ome 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 – Sacchromyces Genome Database, http://www.yeastgenome.org/

References

[1] Schuh RT, Brower AVZ. Biological Systematics: principles and applications. 2nd ed. Cornell University Press; 2009.

[2] Seitz Harald, editor. Analytics of Protein–DNA Interactions. Advances in Biochemical Engineering Biotechnology. Springer; 2007. [3] Pearson H. Genetics: what is a gene? Nature 2006;441(7092):398–401.

[4] UCSC Genome Bioinformatics - <http://hgdownload.cse.ucsc.edu/downloads.html>.

[5] Sims Gregory E, Jun Se-Ran, Wu Guohong A, Kim Sung-Hou. 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 2009;106(8):2677–82.

[6] Murphy William J, Pringle Thomas H, Crider Tess A, Springer Mark S, Miller Webb. Using genomic data to unravel the root of the placental mammal phylogeny. Genome Research 2007;17:413–21.

[7] Zhao Hao, Bourque Guillaume. Recovering genome rearrangements in the mammalian phylogeny. Genome Research 2009;19:934–42.

[8] Prasad Arjun B, Allard Marc W. Confirming the Phylogeny of Mammals by Use of Large Comparative Sequence Data Sets. Molecular Biology and Evolution 2008;25(9):1795–808.

[9] Ebersberger Ingo, Galgoczy Petra, Taudien Stefan, Taenzer Simone, Platzer Matthias, Haeseler Arndt von. Mapping human genetic ancestry. Molecular Biology and Evolution. 2007;24(10):2266–76.

[10] Dunn Casey W et al. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 2008;452:745–50.

[11] Tenreiro Machado JA, Costa António C, Quelhas Maria Dulce. Fractional dynamics in DNA. Communications in Nonlinear Science and Numerical Simulations 2011;16(8):2963–9.

[12] Costa António C, Tenreiro Machado JA. Maria Dulce Quelhas, histogram-based DNA analysis for the visualization of chromosome, genome and species information. Bioinformatics, Vol. 27. Oxford University Press; 2011. 9.

[13] Tenreiro Machado JA, Costa António C, Quelhas Maria Dulce. Entropy analysis of DNA Code dynamics in human chromosomes. Computers and Mathematics with Applications 2011;62(3):1612–7.

[14] Kimura Motoo. The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press; 1983. ISBN 0-521-23109-4.

[15] Deschavanne Patrick J, Giron Alain, Vilain Joseph, Fagot Guillaume, Fertil Bernard. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Molecular Biology and Evolution 1999;16(10):1391–9.

[16] Lynch Michael. The frailty of adaptive hypotheses for the origins of organismal complexity. Proceedings of the National Academy of Sciences of the United States of America 2007;104:8597–604.

[17] Albrecht-Buehler Guenter. Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. Proceedings of the National

Academy of Sciences 2006;103(47):17828–33.

[18] Mitchell David, Bridge Robert. A test of Chargaff's second rule. Biochemical and Biophysical Research Communications 2006;340(1):90–4.

[19] Powdel BR, Satapathy Siddhartha Sankar, Kumar Aditya, Jha Pankaj Kumar, Buragohain Alak Kumar, Borah Munindra, et al. A study in entire chromosomes of violations of the intra-strand parity of complementary nucleotides (Chargaff's second parity rule). DNA Research 2009;16(6):325–43.

[20] Zhang Chun-Ting, Zhang Ren, Ou Hong-Yu. The z curve database: a graphic representation of genome sequences. Bioinformatics 2003;19(5):593–9. [21] Berry MV. Diffractals. Journal of Physics 1979;A12:781–97.

[22] Lapidus ML, Fleckinger-Pellé J. Tambour fractal: vers une résolution de la conjecture de Weyl-Berry pour les valeurs propres du laplacien.

Computational Rend Academy of Sciences Paris Math Sér 1
1988;306:171–5.

[23] Schroeder M. Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise. New York: W.H. Freeman; 1991.