ON THE DNA OF ELEVEN MAMMALS

J. TENREIRO MACHADO
ANTÓNIO C. COSTA
MARIA DULCE QUELHAS

ABSTRACT

This paper studies the DNA code of eleven mammals from the perspective of fractional dynamics. The application of Fourier transform and power law trendlines leads to a categorical representation of species and chromosomes. The DNA information reveals long range memory characteristics.

Keywords: Fractional calculus; Fourier analysis; DNA.

1. Introduction

Fractional calculus (FC) generalizes standard integrals and derivatives to noninteger or even to complex order [Diethelm, 2010; Kilbas *et al.*, 2006; Klimek, 2009; Miller & Ross, 1993; Oldham & Spanier, 1974; Podlubny, 1999; Samko *et al.*, 1993]. During the last decade, it was found that FC plays a fundamental role in modeling of a considerable number of phenomena characterized by long range memory properties [Hilfer, 2000; Machado *et al.*, 2011a; Magin, 2006; Mainardi, 2010; Monje *et al.*, 2010; Oustaloup, 1991; Sabatier *et al.*, 2007; Zaslavsky, 2005]. In fact, FC emerged as the key concept for the study of dynamical systems where classical tools reveal strong limitations. Furthermore, we verify presently that the application of FC concepts encompass a wide spectrum of fields going from physics [Baleanu *et al.*, 2010] and engineering [Lu & Chen, 2010] up to finance [Scalas *et al.*, 2000] and biology [Anastasio, 1994; Ionescu *et al.*, 2011]. This paper studies the deoxyribonucleic acid (DNA) code [Afreixo *et al.*, 2004a; Emanuele II *et al.*, 2005; Jeng *et al.*, 2006; Sims *et al.*, 2009; Yin & Yau, 2005] from the perspective of frac- tional dynamics. It is believed that, for a given living being, besides information about its "struc- tural construction", DNA also includes other levels of information such as the history of its evolution up

to the present state, or instructions for the behavior of each individual during its lifetime [Afreixo *et al.*, 2004b; Dunn *et al.*, 2008; Leit˜ao *et al.*, 2005]. These distinct time scales reveal that we are in the presence of a complex code and that usual tools for the study of dynamical systems may be helpful in this endeavour. This observation motivated the association of logical and mathematical concepts, namely Fourier transform and FC. This work analyzes the DNA data of eleven mammals.

Bearing these ideas in mind, this paper is organized as follows. Section 2 presents the main biological concepts and formulates the framework of the DNA code analysis. Section 3 analyzes the relationship between chromosomes and species for a set of eleven mammals. Finally, Sec. 4 outlines the main conclusions.

2. On the DNA Decoding

DNA is made up of two polymers connected by hydrogen atoms and forming a double helix [Arniker & Kwan, 2009; Pearson, 1999]. Each polymer contains four different nitrogenous bases, namely thymine, cytosine, adenine, and guanine, represented as "T", "C", "A", and "G". Each base on one side bonds with just one type of base on the other side, forming the so-called "base pairing", that is, forming the groups A–T and C–G. For example, in the human being, it was observed that any cell holds 23 pairs of separate DNA-protein complexes (chromosomes), each containing an average of 160 million nucleotide pairs. This massive amount of information is being collected during the past years, as the result of a collaborative effort among many research institutions, and is available for scientific research.

For processing the DNA information we need to start by converting the DNA code into a numerical value. We observe that we are handling an alphabet with four symbols *{*T, C, A, G*}*. In fact, the available data includes a fifth symbol, represented by "N", which was considered by DNA researchers to have no practical meaning for the decoding; therefore, in the sequel this symbol is considered as "zero" during the numerical calculations. The conversion of the four symbols to numerical values must be careful in order to prevent, from inception, any improper effect that may pervade the rest of the numerical treatment. In a previous paper [Machado *et al.*, 2011b], the adoption of Gray code and multiple length sequences was considered. In this paper a simpler process corresponding to direct symbol translation:

$$A = 1 + i0, \quad C = -1 + i0, \quad T = 0 + i, \quad G = 0 - i, \quad N = 0 + i0 \qquad (1)$$

where $i = \sqrt{-1}$, is investigated.

The code assigned in (1) preserves the "base pairing", that is, we have $A = -C$, $T = -G$ and A "orthogonal" to T . This translation scheme is not unique, but is particularly suited for numerical evaluation because it simplifies considerably the rest of the calculations.

Table 1. Main characteristics of the mammals, their tags and chromosomes.

Species	Number	Chromosomes
Human	24	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, Y
Chimpanzee	25	1, 2a, 2b, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, Y
Orangutan	24	1, 2a, 2b, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X
Rhesus	21	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, X
Pig	19	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, X
Opossum	9	1, 2, 3, 4, 5, 6, 7, 8, X
Mouse	21	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, X, Y
Rat	21	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, X
Dog	38	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38
Cow	30	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, X
Horse	32	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, X

Having defined the translation scheme that generates numerical values, the second logical step is to consider that we shall move along the DNA strip, one symbol (base) at a time, and that the resulting values produce a "signal" $x(t)$, t being denoted as the "time" with a restricted meaning. It must be emphasized that we are not referring to any value with units of seconds, but freely describing solely the consecutive base sequencing in the DNA code.

The third processing phase consists of evaluating the characteristics of $x(t)$ in the viewpoint of signal processing and dynamical systems analysis tools [Dodin *et al.*, 2000; Tiwari *et al.*, 1997; Yin & Yau, 2008; Zhou *et al.*, 2007]. In this paper we shall consider the Fourier transform:

$$F\{x(t)\} = X(j\omega) = \int_{-\infty}^{+\infty} x(t)e^{-i\omega t}dt \qquad (2)$$

where F represents the Fourier operator and ω will be named as "angular frequency".

We decided to analyze eleven mammals [Ebersberger *et al.*, 2007; Murphy *et al.*, 2007; Prasad & Bourque, 2008; Zhao & Bourque, 2009] namely, Human, common Chimpanzee, Orangutan, Rhesus monkey, Pig, Opossum, Mouse, Rat, Dog, Cow, and Horse. The chromosomes characteristics of each

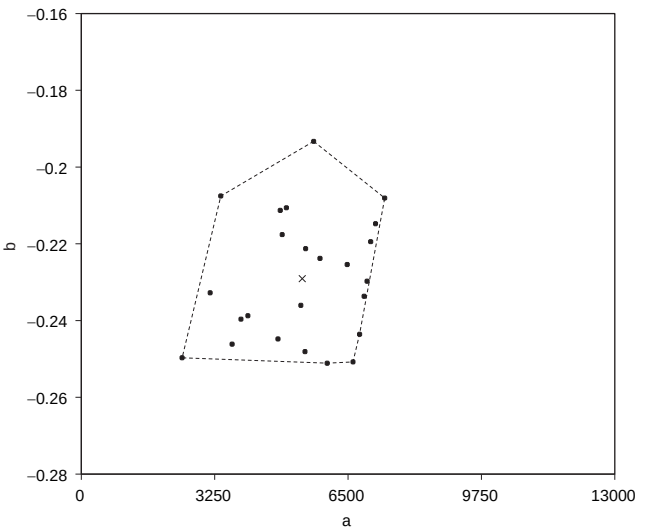
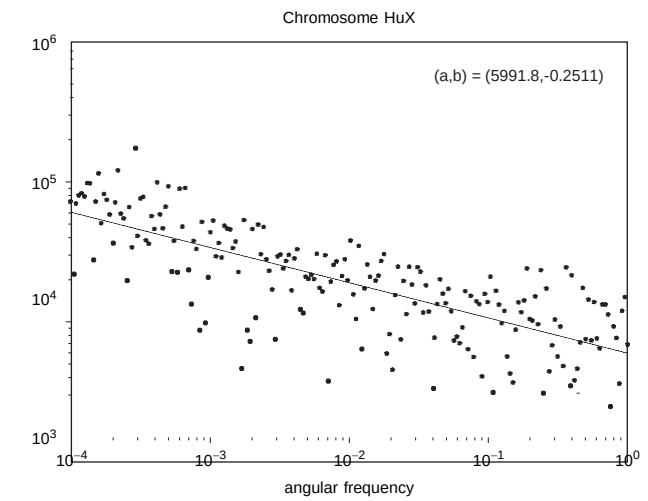


Fig. 2. Locus of the parameters (a, b) for the 24 chromosomes of the Human.



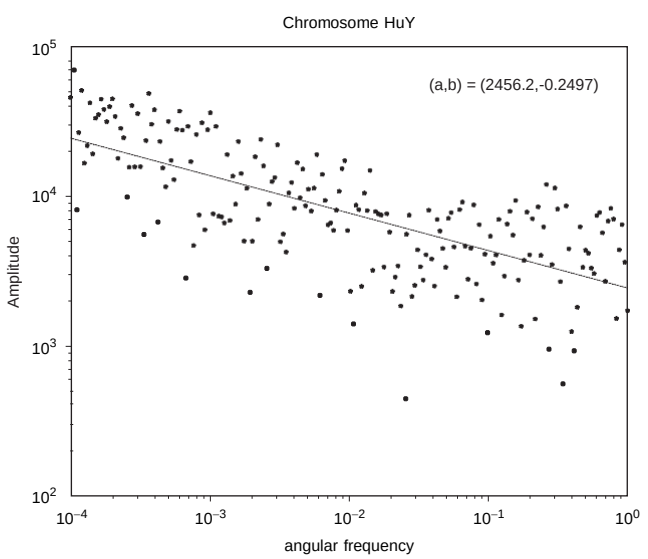


Fig. 1. Fourier transform of the signal for the Human chromosomes X and Y and the corresponding power law approximation.

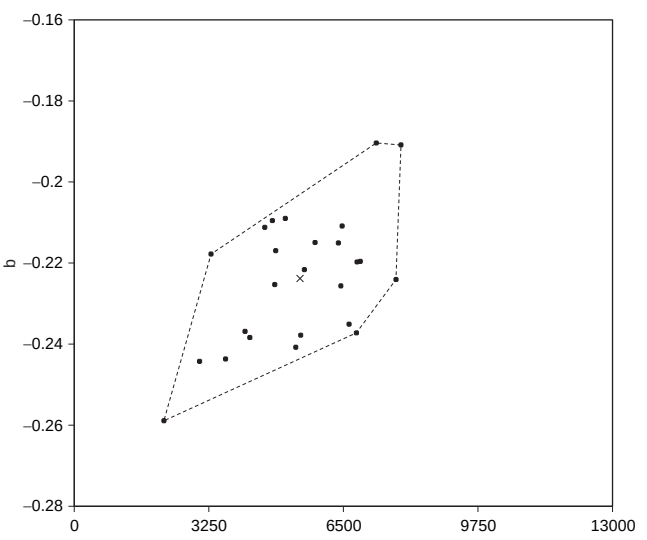


Fig. 3. Locus of the parameters (a, b) for the 25 chromosomes of the Chimpanzee.

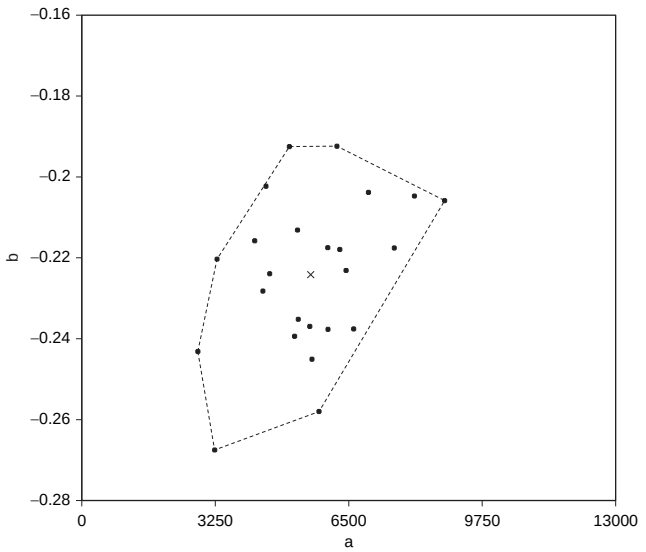


Fig. 4. Locus of the parameters (*a, b*) for the 24 chromosomes of the Orangutan.

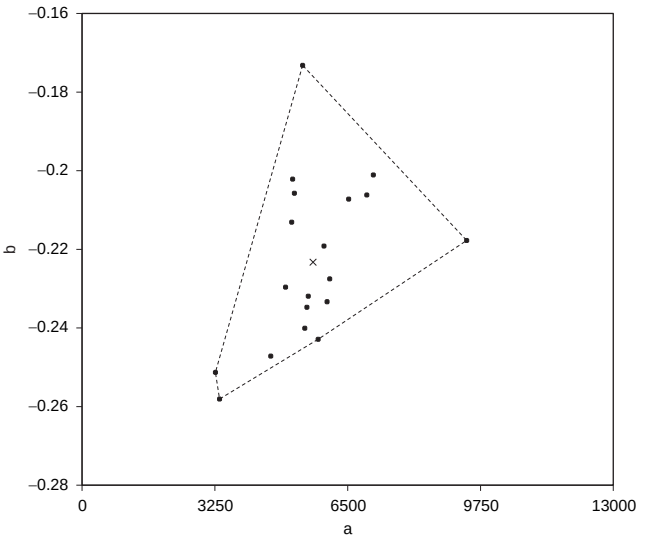


Fig. 6. Locus of the parameters (*a, b*) for the 19 chromosomes of the Pig.

DNA species are presented in Table 1. It should be noted that there is presently no theory, or even an empirical understanding, about the number of chromosomes or their length. The chromosome numbering simply follows a naive classification by size, chromosome 1 being the largest one.

3. Fractional Phenomena

In Sec. 2, a set of eleven mammals was established making up to a total of 265 chromosomes. Therefore, according to the logical reasoning formulated previously, for each chromosome, a "signal" $x_i(t)$, $i = 1, \ldots, 265$, in the perspective of (1) was derived.

The charts of the Fourier transform amplitude reveal that the plot can be approximated by a power function:

$$|F\{x_i(t)\}| \approx a_i \omega^{b_i}, \qquad (3)$$

where $a_i > 0$ and b_i are parameters to be determined by a least square fit procedure. The upper frequency limit is related to the Nyquist sampling theorem, while the lower frequency limit is related

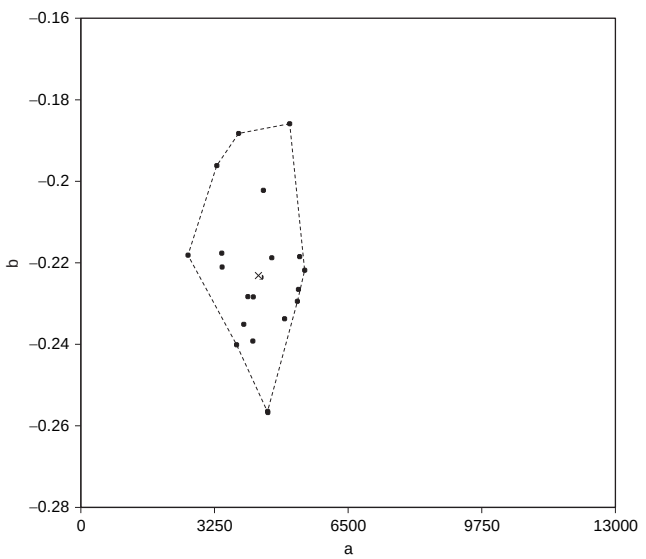


Fig. 5. Locus of the parameters (*a, b*) for the 21 chromosomes of the Rhesus.

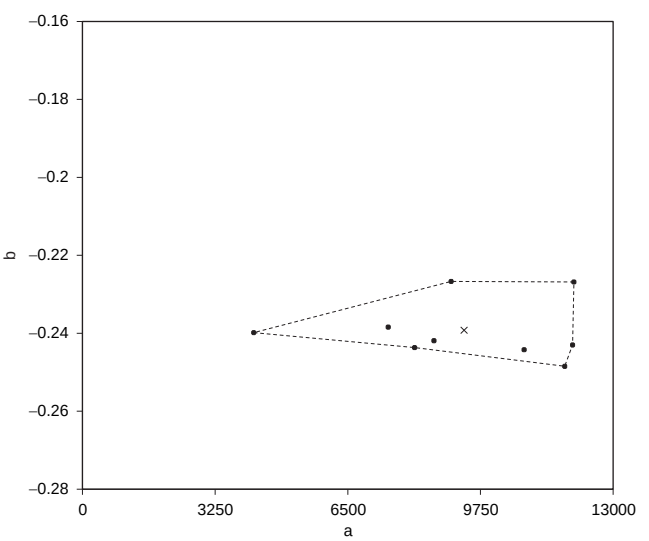


Fig. 7. Locus of the parameters (*a, b*) for the nine chromosomes of the Opossum.

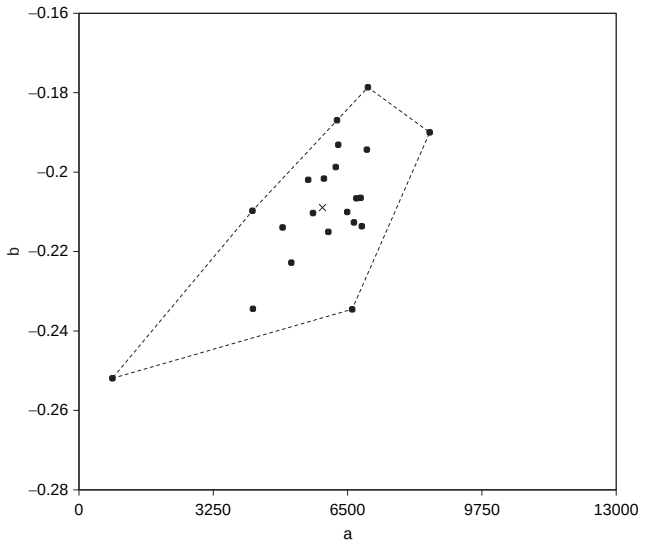


Fig. 8. Locus of the parameters (*a, b*) for the 21 chromosomes of the Mouse.

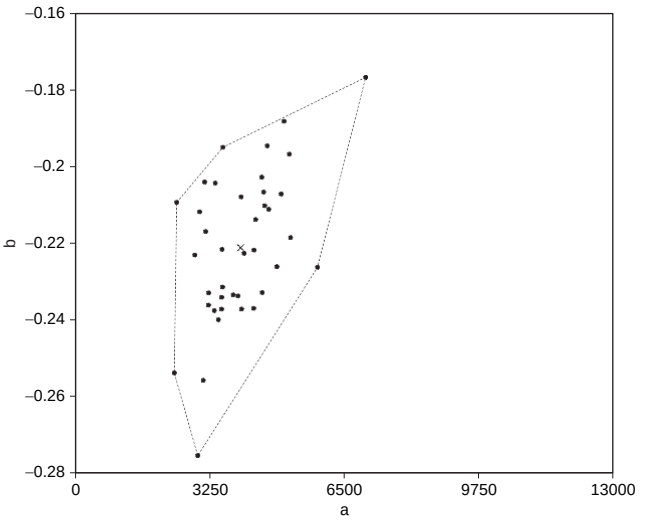


Fig. 10. Locus of the parameters (*a, b*) for the 38 chromosomes of the Dog.

to the total signal length. In both cases, many plots were tested in order to have a reliable bandwidth for the approximation. For example, Fig. 1 depicts the amplitude and the power law approximation for chromosomes HuX and HuY.

Figures 2 to 12 show the locus of parameters (*a*, *b*) for the 265 chromosomes aligned for the eleven mammals. For easing the comparison, the scales are identical in all figures. The cross "x" represents the center of the chromosome set of each species and its coordinates were calculated by averaging the values of the parameters *a* and *b*. The outer contour of the set of chromosomes is also represented for each species. The parameter *a* reflects the size of the chromosome, while parameter *b* describes the nature of the signal information. Therefore, we observe a tendency for smaller/larger values of the point labels to be in the right/left of the locus (*a*, *b*). Furthermore, it is clear that the fractional order of the values is obtained for *b*.

In order to compare the different species, Figs. 13 and 14 depict the superposition of the contours and the centers of the polygons of the eleven mammals, respectively.

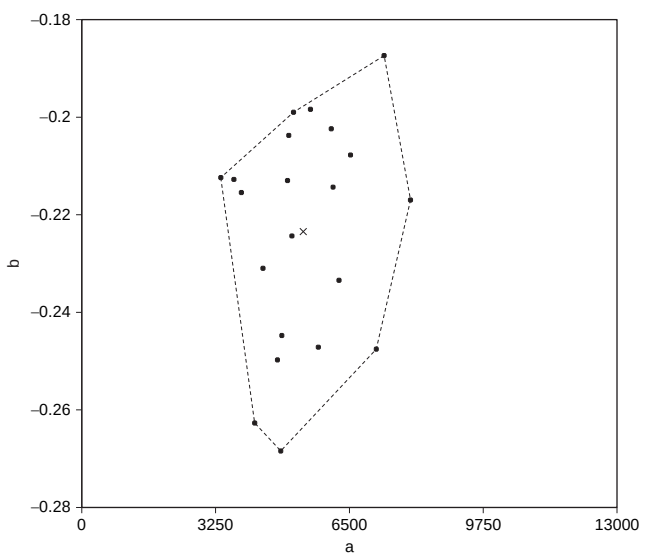


Fig. 9. Locus of the parameters (*a, b*) for the 21 chromosomes of the Rat.

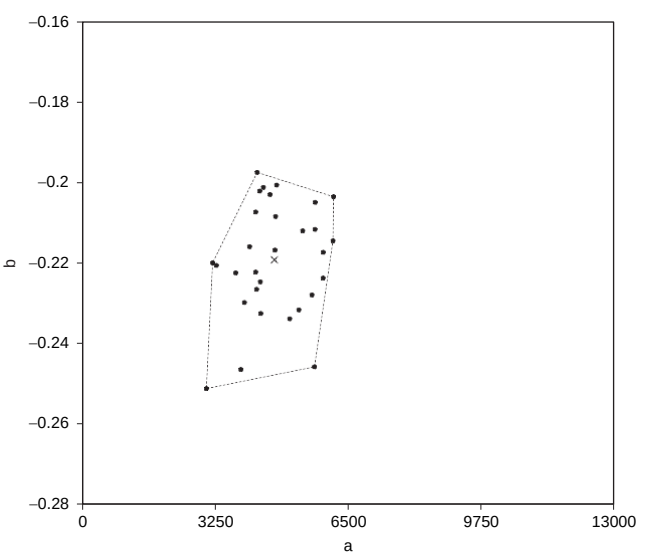


Fig. 11. Locus of the parameters (*a, b*) for the 30 chromosomes of the Cow.

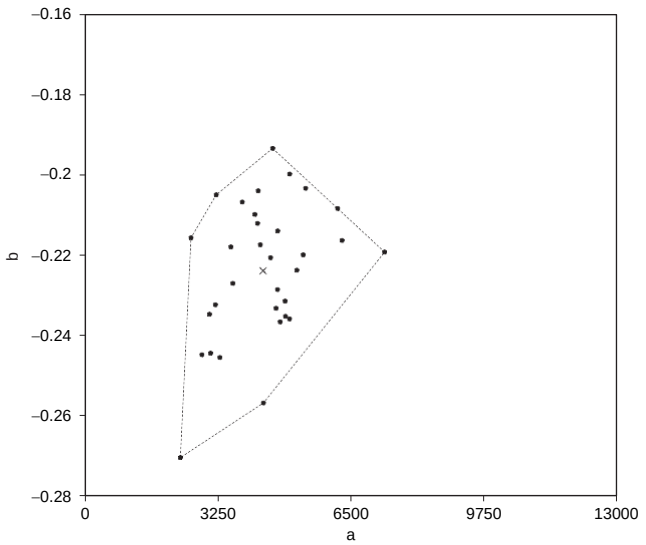


Fig. 12. Locus of the parameters (*a, b*) for the 32 chromosomes of the Horse.

We verify that the Opossum is very different from the rest of the mammals. The Opossum is a marsupial that, in phylogenetic terms, is very far from placental mammals such as the Human or the Horse. Therefore, the result is consistent with phylogenetics. Moreover, we observe that the Chimpanzee and the Orangutan are close to the Human. Although slightly apart, the Pig and the Rat are also relatively close to the Human. On the other hand, the Rhesus and the Horse reveal strong similarities between themselves. Therefore, most of these results not only follow what is known qualitatively, but also provide a quantitative tool for

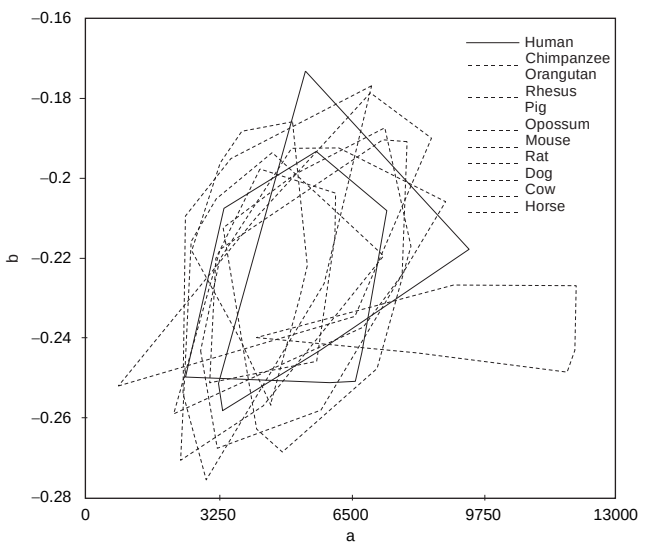


Fig. 13. Locus (*a, b*) of contour of the eleven mammals.

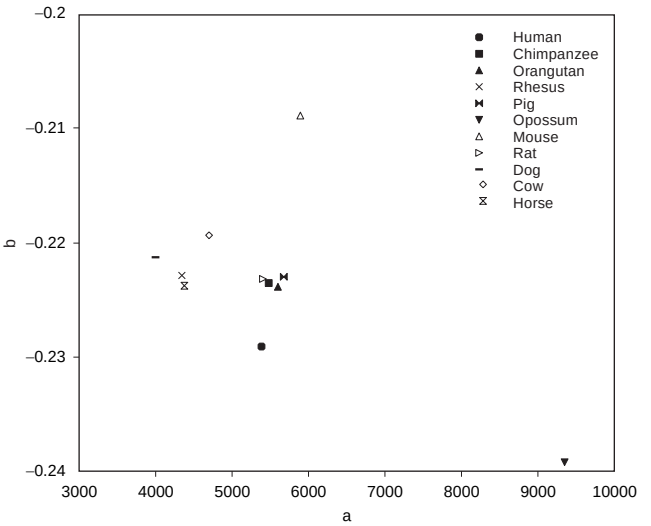


Fig. 14. Locus (*a, b*) of centers of polygons of the eleven mammals.

measuring "distances" between species in the viewpoint of phylogenetics. At a different level, the results demonstrate that the information coded in the DNA has long memory and that characteristic is easily captured by the tools of FC.

4. Conclusions

In this paper, the chromosomes four symbol alphabet was converted to numerical values and the resulting signal, representative of the DNA code, was processed by means of the Fourier transform. The amplitude of the Fourier transform was characterized by power law approximations typical of systems with fractional dynamics. The locus of power law parameters revealed important details, namely clusters of species and chromosomes. The results are in agreement with what is currently known in phylogenetics and motivate further research efforts in the perspective of FC.

Acknowledgments

The authors 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.cgi GENOME=Pongo

- Rhesus — Macaque Genome Sequencing Consortium, http://www.hgsc.bcm.tmc.edu/projects/rmacaque/
- Pig — The Swine Genome Sequencing Consortium, http://piggenome.org/
- Cow — The Baylor College of Medicine Human Genome Sequencing Center, http://www.hgsc.bcm.tmc.edu/projects/bovine/
- Dog — Genome Sequencing Project http://www.broad.mit.edu/mammals/dog/, Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**(Dec 8), 803–819 (2005).
- Horse — The Broad Institute, http://www.broad.mit.edu/mammals/horse/
- Mouse — Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002), http://www.hgsc.bcm.tmc.edu/projects/mouse/
- Rat — The Baylor College of Medicine Human Genome Sequencing Center, http://www.hgsc.bcm.tmc.edu/projects/rat/,
  Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**(6982), 493–521 (2004)
- Opossum — The Broad Institute, http://www.broad.mit.edu/mammals/opossum/
- UCSC Genome Bioinformatics http://hgdownload.cse.ucsc.edu/downloads.html

References

Afreixo, V., Ferreira, P. & Santos, D. [2004a] "Fourier analysis of symbolic data: A brief review," *Dig. Sign. Process.* **14**, 523–530.

Afreixo, V., Ferreira, P. & Santos, D. [2004b] "Spectrum and symbol distribution of nucleotide sequences," *Phys. Rev. E* **70**, 031910-1–4.

Anastasio, T. [1994] "The fractional-order dynamics of brainstem vestibulo-oculomotor neurons," *Biol. Cybern.* **72**, 69–79.

Arniker, S. & Kwan, H. [2009] "Graphical representation of DNA sequences," *Proc. 2009 IEEE Int. Conf. Electro/Information Technology*, pp. 311–314.

Baleanu, D., Golmankhaneh, A., Golmankhaneh, A. & Nigmatullin, R. [2010] "Newtonian law with memory," *Nonlin. Dyn.* **60**, 81–86.

Diethelm, K. [2010] *The Analysis of Fractional Differential Equations* (Springer, Berlin).

Dodin, G., Vandergheynst, P., Levoir, P., Cordier, C. & Marcourt, L. [2000] "Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences," *J. Theoret. Biol.* **206**, 323–326.

Dunn, C. *et al.* [2008] "Broad phylogenomic sampling improves resolution of the animal tree of life," *Nature* **452**, 745–749.

Ebersberger, I., Galgoczy, P., Taudien, S., Taenzer, S., Platzer, M. & von Haeseler, A. [2007] "Mapping human genetic ancestry," *Mol. Biol. Evol.* **24**, 2266–2276.

Emanuele II, V., Tran, T. & Zhou, G. [2005] "A Fourier product method for detecting approximate TANDEM repeats in DNA," *Proc. 2005 IEEE/SP 13th Workshop on Statistical Signal Processing*, pp. 1390–1395.

Hilfer, R. [2000] *Applications of Fractional Calculus in Physics* (World Scientific Publishing Company, Singapore).

Ionescu, C., Machado, J. & De Keyser, R. [2011] "Modeling of the lung impedance using a fractional order ladder network with constant phase elements," *IEEE Trans. Biomed. Circuits Syst.* **5**, 83–89.

Jeng, C., Yang, Y., Hsieh, K. & Lin, C. [2006] "Clustering analysis for bacillus genus using Fourier transform and self-organizing map," *ICONIP 2006*, Part III, LNCS, Vol. 4234 (Springer-Verlag), pp. 48–57.

Kilbas, A., Srivastava, H. & Trujillo, J. [2006] *Theory and Applications of Fractional Differential Equations* (Elsevier, Amsterdam).

Klimek, M. [2009] *On Solutions of Linear Fractional Differential Equations of a Variational Type* (Czestochowa University of Technology, Czestochowa).

Leita˜o, H., Pessoa, L. & Stolfi, J. [2005] "Mutual information content of homologous DNA sequences," *Genet. Mol. Res.* **4**, 553–562.

Lu, G. & Chen, Y. [2010] "Robust stability and stabilization of fractional-order interval systems with the fractional order alpha: The $0 < \alpha < 1$ case," *IEEE Trans. Autom. Contr.* **55**, 152–158.

Machado, J., Kiryakova, V. & Mainardi, F. [2011a] "Recent history of fractional calculus," *Commun. Nonlin. Sci. Numer. Simul.* **16**, 1140–1153.

Machado, J., Costa, A. & Quelhas, M. [2011b] "Fractional dynamics in DNA," *Commun. Nonlin. Sci. Numer. Simul.* **16**, 2963–2969.

Magin, R. [2006] *Fractional Calculus in Bioengineering* (Begell House Publishers, Redding).

Mainardi, F. [2010] *Fractional Calculus and Waves in Linear Viscoelasticity: An Introduction to Mathematical Models* (Imperial College Press, London).

Miller, K. & Ross, B. [1993] *An Introduction to the Fractional Calculus and Fractional Differential Equations* (John Wiley & Sons, NY).

Monje, C., Chen, Y., Vinagre, B., Xue, D. & Feliu, V. [2010] *Fractional Order Systems and Controls*: *Fundamentals and Applications* (Springer, London).

Murphy, W., Pringle, T., Crider, T., Springer, M. & Miller, W. [2007] "Using genomic data to unravel the root of the placental mammal phylogeny," *Gen. Res.* **17**, 413–421.

Oldham, K. & Spanier, J. [1974] *The Fractional Calculus*: *Theory and Application of Differentiation and Integration to Arbitrary Order* (Academic Press, NY and London).

Oustaloup, A. [1991] *La Commande CRONE* : *Commande Robuste dOrdre Non Entier* (Herm`es, Paris).

Pearson, H. [1999] "Genetics: What is a gene?" *Nature* **441**, 398–401.

Podlubny, I. [1999] *Fractional Differential Equations* (Academic Press, San Diego).

Prasad, A. & Allard, M. [2008] "Confirming the phylogeny of mammals by use of large comparative sequence data sets," *Mol. Biol. Evol.* **25**, 1795–1808.

Sabatier, J., Agrawal, O. & Machado, J. [2007] *Advances in Fractional Calculus. Theoretical Developments and Applications in Physics and Engineering* (Springer, Dordrecht).

Samko, S., Kilbas, A. & Marichev, O. [1993] *Fractional Integrals and Derivatives*: *Theory and Applications*, 2nd edition (Gordon & Breach Science Publishers, London).

Scalas, E., Gorenflo, R. & Mainardi, F. [2000] "Fractional calculus and continuous-time finance," *Physica A* **284**, 376–384.

Sims, G., Jun, S., Wu, G. & Kim, S. [2009] "Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions," *Proc. Nat. Acad. Sci. USA* **106**, 2677–2682.

Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S. & Ramaswamy, R. [1997] "Prediction of probable genes by Fourier analysis of genomic sequences," *Comput. Appl. Biosci.* **13**, 263–270.

Yin, C. & Yau, S. [2005] "A Fourier characteristic of coding sequences: Origins and a non-Fourier approximation," *J. Comput. Biol.* **12**, 1153–1165.

Yin, C. & Yau, S. [2008] "Numerical representation of DNA sequences based on genetic code context and its applications in periodicity analysis of genomes," *IEEE Symp. Computational Intelligence in Bioinformatics and Computational Biology*, pp. 223–227.

Zaslavsky, G. [2005] *Hamiltonian Chaos and Fractional Dynamics* (Oxford University Press, NY).

Zhao, H. & Bourque, G. [2009] "Recovering genome rearrangements in the mammalian phylogeny," *Gen. Res.* **19**, 934–942.

Zhou, Y., Zhou, L., Yu, Z. & Anh, V. [2007] "Distinguish coding and noncoding sequences in a complete genome using Fourier transform," *Proc. IEEE Third Int. Conf. Natural Computation*, pp. 295–299.