Computers and Mathematics with Applications 62 (2011) 1612–1617

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

Entropy analysis of the DNA code dynamics in human chromosomes

J.A. Tenreiro Machado [a,*], António C. Costa [b,1], Maria Dulce Quelhas [c,2]

[a] Institute of Engineering of Porto, Department of Electrical Engineering, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal
[b] Institute of Engineering of Porto, Department of Informatics Engineering, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal
[c] National Health Institute, Medical Genetics Center, Praça Pedro Nunes, 88, 4099-028, Porto, Portugal

ARTICLE INFO

Keywords:
DNA
Chromosome
Fractional calculus
Entropy
Dynamics

ABSTRACT

Deoxyribonucleic acid, or DNA, is the most fundamental aspect of life but present day scientific knowledge has merely scratched the surface of the problem posed by its decoding. While experimental methods provide insightful clues, the adoption of analysis tools supported by the formalism of mathematics will lead to a systematic and solid build-up of knowledge. This paper studies human DNA from the perspective of system dynamics. By associating entropy and the Fourier transform, several global properties of the code are revealed. The fractional order characteristics emerge as a natural consequence of the information content. These properties constitute a small piece of scientific knowledge that will support further efforts towards the final aim of establishing a comprehensive theory of the phenomena involved in life.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Fractional calculus (FC) deals with the generalization of integrals and derivatives to a non-integer order. FC extends to a wide field of applications by bringing into a broader paradigm concepts of physics, mathematics and engineering [1–13]. From this perspective, FC encompasses all areas of scientific knowledge and recent research has demonstrated the usefulness of FC concepts in the study of a plethora of distinct dynamical phenomena.

Entropy was introduced in thermodynamics by Clausius (1862) and Boltzmann (1896). Later the concept was applied by Shannon [14] and Jaynes [15] to information theory. During the last few years alternative entropy measures have been proposed and have opened up novel directions for the adoption of this concept in complex system analysis [14–27].

Deoxyribonucleic acid (DNA) encodes a huge amount of information responsible for the structure of each living being. This chemical structure masters all aspects involved not only in the body development and functioning during the lifetime of a given individual, but also the evolutionary process in the scope of the interaction of that species with the natural world. The DNA code is, therefore, the most fundamental property of life, but present day human knowledge captures only small glimpses of the formidable quantity of information. Understanding the code implemented by the DNA array is probably one of the most important challenges of human science and, therefore, it is relevant to adopt methods of analysis supported by the solid formalisms of mathematical tools.

With the advent of genome sequencing and genome databases, considerable information has become available for computational processing, motivating research on understanding the information structure of DNA. The present paper

* Corresponding author. Tel.: +351 22 8340500; fax: +351 22 8321159.
E-mail addresses: jtm@isep.ipp.pt (J.A.T. Machado), acc@isep.ipp.pt (A.C. Costa), mdquelhas@gmail.com (M.D. Quelhas).
URLs: http://ave.dee.isep.ipp.pt/~jtm/ (J.A.T. Machado), http://www.linkedin.com/in/amccosta (A.C. Costa).
1 Tel.: +351 22 8340500; fax: +351 22 8321159.
2 Tel.: +351 22 6070306.

0898-1221/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2011.03.005

Table 1
Size of the human chromosomes.

Chromosome	File size with the "N" (kBytes)	File size without the "N" (kBytes)
Hu1	254235	225281
Hu2	248063	238205
Hu3	201982	194797
Hu4	194977	187662
Hu5	184533	177695
Hu6	174537	167395
Hu7	162321	155354
Hu8	149291	142889
Hu9	144038	120143
Hu10	138245	131315
Hu11	137707	131130
Hu12	136529	130481
Hu13	117473	95589.9
Hu14	109497	88289.5
Hu15	104582	81694.8
Hu16	92161.9	78884.7
Hu17	82819.1	77795.2
Hu18	79638.8	74657.2
Hu19	60311.6	55809
Hu20	64286	59505.5
Hu21	49092.5	35106.6
Hu22	52330.7	34894.5
HuX	158376	151101
HuY	60561	25653.6

studies the deoxyribonucleic acid (DNA) human code from the perspective of system dynamics [29–40]. A first approach to the DNA decoding merely requires "static tools", such as statistics; nevertheless "dynamic tools" may prove to be powerful allies in the scientific processing.

These ideas motivated the association of the mathematical concepts of FC, entropy and the Fourier transform for the analysis of the DNA data involved in the human twenty three chromosome pairs, of which twenty four chromosomes are distinct. The results reveal important properties, demonstrating the value of the proposed methodology, and motivating further research with the tools of fractional dynamics. Therefore, the aim of this study is twofold, since it intends not only to take advantage of the association of several tools usual in system dynamical analysis, but also to investigate the characteristics of long range memory effects that are highlighted by FC.

Having these ideas in mind, this paper is organized as follows. Section 2 presents the fundamental biological concepts and the mathematical tools, and formulates their application for the DNA sequence decoding. Section 3 analyzes the information content of human chromosomes from the perspective of FC. Finally, Section 4 outlines the main conclusions.

2. Mathematical tools and DNA decoding

DNA is made up of two polymers connected by the bonding of hydrogen atoms, leading to a double-helix structure [28,29]. Each polymer contains nucleotides that can be classified into three types: deoxyribose (a five-carbon sugar), a phosphate group, and a nitrogenous base. There are four different nitrogenous bases: thymine, cytosine, adenine, and guanine, denoted by the symbols $\{T, C, A, G\}$. Each type of base on one strand forms a bond with just one type of base on the other strand. This arrangement is called "base pairing", with A and C bonding only to T and G, respectively. The four bases are the foundation of the genetic code and instruct cells on how to synthesize enzymes and proteins. In a human being each cell holds twenty three pairs of separate DNA–protein complexes (chromosomes), each containing, on average, 160 million nucleotide pairs. This massive amount of information has been being collected and decoded during the last few years, as the result of a large collaborative effort among many individuals and at research institutions around the world, and is freely available [31] for scientific research.

Due to the vast volume of information, a histogram-based measure was adopted. However, in general, histograms do not capture dynamics. In order to overcome this limitation two approaches were considered, namely the construction of histograms with n-tuple base sequences for bin counting and the adoption of a sliding window. In the first case the algorithm counts sequences of length n composed of the four base symbols. The available chromosome data include a fifth symbol, represented by "N", which has no practical meaning for the DNA decoding. In fact, DNA researchers consider that some parts of the DNA code contain information that is not relevant and marked it with the symbol "N". The authors analyzed the effect of considering, or not considering, this fifth symbol and verified it to be of minor importance. Therefore, this symbol was discarded during the histogram bin construction. We have different statistics when considering length ranging from $n = 1$, representing merely a static counting of $m = 4^1$ states, up to $n = 3$, representing the dynamics of a system with $m = 4^3$ (64) states. In other words, we define the bins $[\{A\}, \{C\}, \{G\}, \{T\}]$, $[\{AA\}, \{AC\}, \ldots, \{TC\}, \{TA\}]$, and $[\{AAA\}, \{AAG\}, \ldots, \{TAC\}, \{TAA\}]$ for $n = 1$, $n = 2$ and $n = 3$, respectively. A partial overlapping

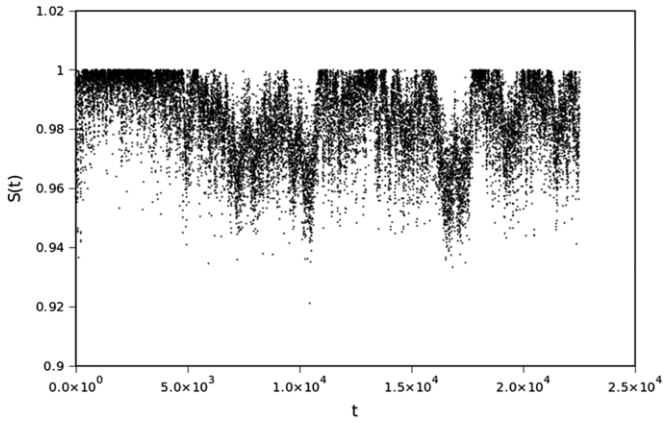

Fig. 1. Entropy variation $S(t)$ for human chromosome Hu1 when $n = 1$ and $W = 10^4$.

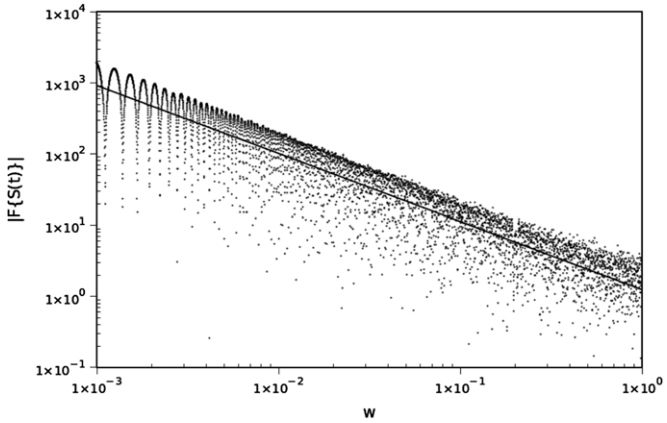

Fig. 2. Amplitude of the Fourier spectrum $|F\{S(t)\}|$ versus ω and the power law trendline approximation for human chromosome Hu1 when $n = 1$ and $W = 10^4$.

of $n - 1$ base sequences is considered, Therefore, for a DNA strand of length L and bins of length n, there result a total of $L - n + 1$ counts. In the second case, for the bin counting, a sliding window of width W is adopted. When counting the n-tuples, once the window has elapsed a new histogram is built and the corresponding statistical measure is evaluated.

Once the histograms are collected, the second step in our analysis consists in evaluating their characteristics. In this paper we consider the Shannon entropy S defined by [14]

$$S = -\sum_{i=1}^{W} p_i \ln(p_i) \tag{1}$$

where the window width W is the number of possible events and p_i is the probability that event i occurs, so $\sum_{i=1}^{W} p_i = 1$.

We have a sequence of windows along the DNA strand and, therefore, a "signal" $S(t)$ is generated, where t may be interpreted as the "time" progress during each histogram construction. In general $S(t)$ is noisy and difficult to analyze in the "time domain". Therefore, the global dynamical characteristics of $S(t)$ may only emerge on applying signal processing tools [41]. In the present study we adopt the Fourier transform, defined as [42]

$$F\{S(t)\} = \int_{-\infty}^{+\infty} S(t) \mathrm{e}^{-\mathrm{j}\omega t} \mathrm{d}t \tag{2}$$

where ω represents the angular frequency and $\mathrm{j} = \sqrt{-1}$.

It should be noticed that we use the notions of "time" and "angular frequency" with units time and its inverse. This is something of an abuse since t describes the progression along the DNA strand, but is it adopted with care in the absence of a better term.

Another important aspect to discuss is that we are converting a message, with some unknown code, into a numerical signal. All assumptions introduced *a priori* may "deform" the message. Therefore, the methodology adopted tries to be the least invasive possible by adopting tools that capture, process and describe global characteristics.

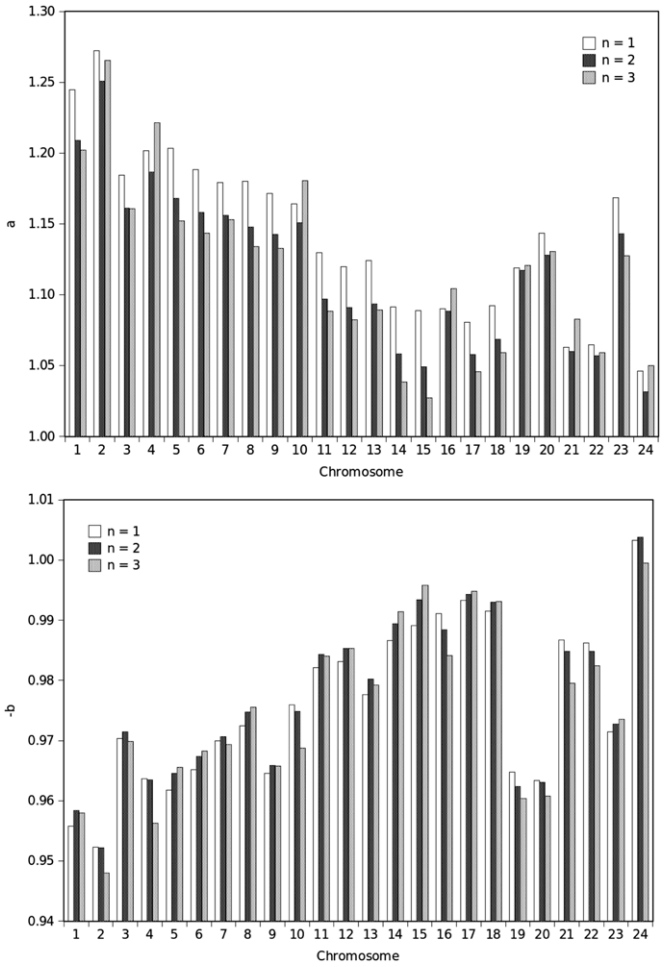

Fig. 3. Power law trendline parameters (a, b) for {Hu1, …, Hu22, HuX, HuY} when $n = \{1, 2, 3\}$, $W = 10^4$.

In synthesis, for the DNA human analysis we adopt (i) histograms for translating the DNA code by counting n-tuple base sequences in windows of width W, (ii) entropy for "signal" $S(t)$ generation and (iii) the Fourier transform for signal analysis. Pursuing this line of thought, in the next section we consider the human DNA and the chromosomes {Hu1, Hu2, …, HuX, HuY}. The characteristics of these twenty four chromosomes, namely the file sizes with and without the "N" symbol, are depicted in Table 1.

3. Fourier and entropy analysis of human DNA

For each of the twenty four distinct human chromosomes the corresponding $S(t)$ was calculated. These signals have a strong variability, making their direct comparison in the time domain difficult. These characteristics are present in all chromosomes and for different values of W and n. When we have calculated the Fourier transform, the spectrum reveals that the characteristics are close to those encountered in fractional systems. For analyzing the dynamics, the power law trendline $|F\{S(t)\}| = a\omega^b$, $a \in \Re^+$, $b \in \Re$, was superimposed for each case, leading to an adequate fit of the numerical data.

For example, Figs. 1 and 2 depict the signal $S(t)$ and the corresponding Fourier amplitude $|F\{S(t)\}|$ versus frequency ω, with the corresponding trendline, for Hu1, when $n = 1$ and $W = 10^4$. We observe clearly signal–noise characteristics, a non-integer value of b and fractional order behavior. In fact both curves reveal a considerable chattering, typical of noisy signals, and a closer look at the plots shows that the "time" curve has a fractal-like behavior, as demonstrated by the "frequency" response, approximately given by $|F\{S(t)\}| = 1.245\omega^{-0.956}$, with a clear non-integer order slope.

Figs. 3 and 4 show power law trendline parameters (a, b) of {Hu1, …, Hu22, HuX, HuY} for $n = \{1, 2, 3\}$, $W = 10^4$ and for $n = 1$, $W = \{10^2, 10^3, 10^4\}$, respectively. Chromosomes X and Y are listed at the end with numbers 23 and 24. It must be noted that W establishes a pseudo-time sampling and, therefore, a Nyquist frequency of $\omega_W = \frac{\pi}{W}$. Therefore, during the Fourier analysis and the trendline calculation the "frequency" bandwidth $\omega \in [\omega_{\min}, \omega_{\max}]$ was adjusted to $\omega_{\min} = \left\{10^{-5}, 10^{-4}, 10^{-3}\right\}$ and $\omega_{\max} = \left\{10^{-2}, 10^{-1}, 10^{0}\right\}$ for $W = \{10^2, 10^3, 10^4\}$.

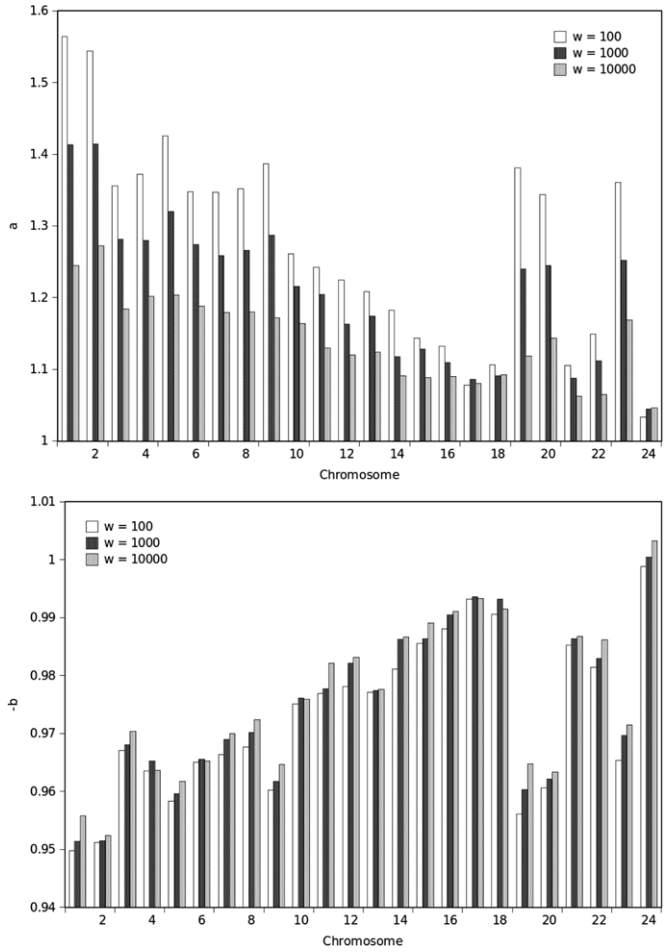

Fig. 4. Power law trendline parameters (a, b) for {Hu1, …, Hu22, HuX, HuY} when $n = 1$, $W = \{10^2, 10^3, 10^4\}$.

We verify that we get slightly different numerical values when adjusting n and W, but that the global fractional order dynamics remains, revealing the long range memory dependence of the DNA encoding. In practical terms, this result should be expected due to two main signal characteristics, namely its fractal structure and its length. Both aspects were already mentioned in relation to Fig. 1 and Table 1, but even so it is important to realize that the signal truncation, often adopted when processing common signals, eliminates implicitly long range effects. In the present case, the analysis tool encompasses a considerable history and, therefore, allows the emergence of fractional dynamics.

Another interesting aspect emerges from comparing the fractional order characteristics of the distinct chromosomes. It should be noted that the chromosome numbering follows the naïf historical logic of simply following their size, with 1 for the largest and 22 for the smallest, with the two sexual chromosomes designated as X (the larger) and Y (the smaller). It is easily noticeable that a clear pattern emerges, namely with smaller values for larger chromosome number, with the parameter a following the same type of behavior as b. Chromosomes X and Y, listed at the end, have sizes between Chr7 and Chr8, and Chr20 and Chr21, respectively, as can be seen in Table 1.

While parameter a is simply correlated with the chromosome size and the total volume of information, b addresses a more subtle aspect, namely the signal information content. Since the closer the value of b is to zero the larger the randomness of $S(t)$ is, we conclude that the size logic lies, in fact, beneath the logic of stronger information variation. By other words, the larger the chromosome, the greater the scope for variation of the information content.

In conclusion, the combination of mathematical tools adopted has revealed fractional dynamics of the DNA code typical of long range memory phenomena, and suggests the pursuance of this line of research for disclosing DNA information.

4. Conclusions

Chromosomes have a code based on a four-symbol alphabet that programs all aspects of life and evolution. Decoding of DNA is taking its first steps and both experimental and mathematical research strategies constitute valid approaches for extracting pieces of the puzzle presented by nature. This information can be analyzed with tools usually adopted in the study of dynamical systems. A quantitative analysis must avoid introducing assumptions that may distort all subsequent numerical

processing. The proposed methodology embedding histograms, entropy and the Fourier transform led to the emergence of fractional dynamics and to a straightforward interpretation of the results. It is believed that the weak restrictions assumed in the quantifying phase have a reduced impact upon the global message dynamics. In other words, it is our understanding that the measuring apparatus has a minor influence upon the measured phenomenon. The results reflect global characteristics that should remain invariant under distinct methods. Therefore, other strategies of attacking the code, namely adopting tools with a more detailed nature, can be explored, having in mind global properties that should be preserved.

Acknowledgements

We thank the following organizations for allowing access to genome data of the Human-Genome Reference Consortium: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/.

References

[1] Keith B. Oldham, Jerome Spanier, The Fractional Calculus: Theory and Application of Differentiation and Integration to Arbitrary Order, Academic Press, 1974.
[2] Stefan G. Samko, Anatoly A. Kilbas, Oleg I. Marichev, Fractional Integrals and Derivatives: Theory and Applications, Gordon and Breach Science Publishers, 1993.
[3] Kenneth S. Miller, Bertram Ross, An Introduction to the Fractional Calculus and Fractional Differential Equations, John Wiley and Sons, 1993.
[4] Alain Oustaloup, La Commande CRONE: Commande Robuste d'ordre Non Entier, Hermes, 1991.
[5] Igor Podlubny, Fractional Differential Equations, Academic Press, San Diego, 1999.
[6] A.A. Kilbas, H.M. Srivastava, J.J. Trujillo, Theory and Applications of Fractional Differential Equations, in: North-Holland Mathematics Studies, vol. 204, Elsevier, 2006.
[7] Richard L. Magin, Fractional Calculus in Bioengineering, Begell House Publishers, 2006.
[8] J. Sabatier, O.P. Agrawal, J.A. Tenreiro Machado, Advances in Fractional Calculus: Theoretical Developments and Applications in Physics and Engineering, Springer, 2007.
[9] Francesco Mainardi, Fractional Calculus and Waves in Linear Viscoelasticity: An Introduction to Mathematical Models, Imperial College Press, 2010.
[10] Riccardo Caponetto, Giovanni Dongola, Luigi Fortuna, Ivo Petráš, Fractional Order Systems: Modeling and Control Applications, World Scientific Publishing Company, 2010.
[11] Concepcion Alicia Monje, YangQuan Chen, Blas Manuel Vinagre, Dingyu Xue, Vicente Feliu, Fractional Order Systems and Controls: Fundamentals and Applications, in: Series: Advances in Industrial Control, Springer, 2010.
[12] Kai Diethelm, The Analysis of Fractional Differential Equations, in: Lecture Notes in Mathematics Series, Springer, 2010.
[13] J.Tenreiro Machado, Virginia Kiryakova, Francesco Mainardi, Recent History of Fractional Calculus, in: Communications in Nonlinear Science and Numerical Simulations, 16, Elsevier, 2011, pp. 1140–1153.
[14] C.E. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (1948) 379–423. 623–656.
[15] E.T. Jaynes, Information theory and statistical mechanics, Physical Review 106 (1957) 620.
[16] A.I. Khinchin, Mathematical Foundations of Information Theory, Dover, New York, 1957.
[17] A. Plastino, A.R. Plastino, Tsallis entropy and Jaynes information theory formalism, Brazilian Journal of Physics 29 (1) (1999) 50–60.
[18] X. Li, C. Essex, M. Davison, K.H. Hoffmann, C. Schulzky, Fractional diffusion, irreversibility and entropy, Journal of Non-Equilibrium Thermodynamics 28 (3) (2003) 279–291.
[19] H.J. Haubold, A.M. Mathai, R.K. Saxena, Boltzmann–Gibbs entropy versus Tsallis entropy: recent contributions to resolving the argument of Einstein concerning neither Herr Boltzmann nor Herr Planck has given a definition of W? Astrophysics and Space Science 290 (3–4) (2004) 241–245.
[20] A.M. Mathai, H.J. Haubold, Pathway model, superstatistics, Tsallis statistics, and a generalized measure of entropy, Physica A: Statistical Mechanics and its Applications 375 (1) (2007) 110–122.
[21] T. Carter, An Introduction to Information Theory and Entropy, Complex Systems Summer School, Santa Fe, 2007.
[22] P. Rathie, S. da Silva, Shannon, Levy, and Tsallis: a note, Applied Mathematical Science 2 (28) (2008) 1359–1363.
[23] C. Beck, Generalised information and entropy measures in physics, Contemporary Physics 50 (4) (2009) 495–510.
[24] R.M. Gray, Entropy and Information Theory, Springer-Verlag, 2009.
[25] M.R. Ubriaco, Entropies based on fractional calculus, Physics Letters A 373 (30) (2009) 2516–2519.
[26] I.G. Taneja, On measures of information and inaccuracy, Journal of Statistical Physics 14 (1976) 203–270.
[27] B.D. Sharma, R.K. Taneja, Three generalized additive measures of entropy, Elektronische Informationsverarbeitung und Kybernetik (EIK) 13 (1977) 419–433.
[28] A. Wehrl, General properties of entropy, Reviews of Modern Physics 50 (1978) 221–260.
[29] David Micklos, Greg A. Freyer, DNA Science: A First Course, 2nd ed., Cold Spring Harbor Laboratory Press, 2003.
[30] James D. Watson, DNA: The Secret of Life, Arrow Books Ltd, 2004.
[31] R.T. Schuh, A.V.Z. Brower, Biological Systematics: Principles and Applications, 2nd ed., Cornell University Press, 2009.
[32] Harald Seitz (Ed.), Analytics of Protein–DNA Interactions, in: Advances in Biochemical Engineering Biotechnology, Springer, 2007.
[33] H. Pearson, Genetics: what is a gene? Nature 441 (7092) (2006) 398–401.
[34] UCSC Genome Bioinformatics. http://hgdownload.cse.ucsc.edu/downloads.html.
[35] Gregory E. Sims, Se-Ran Jun, Guohong A. Wu, Sung-Hou 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 106 (8) (2009) 2677–2682.
[36] William J. Murphy, Thomas H. Pringle, Tess A. Crider, Mark S. Springer, Webb Miller, Using genomic data to unravel the root of the placental mammal phylogeny, Genome Research 17 (2007) 413–421.
[37] Hao Zhao, Guillaume Bourque, Recovering genome rearrangements in the mammalian phylogeny, Genome Research 19 (2009) 934–942.
[38] Arjun B. Prasad, Marc W. Allard, Confirming the phylogeny of mammals by use of large comparative sequence data sets, Molecular Biology and Evolution 25 (9) (2008) 1795–1808.
[39] Ingo Ebersberger, Petra Galgoczy, Stefan Taudien, Simone Taenzer, Matthias Platzer, Arndt von Haeseler, Mapping human genetic ancestry, Molecular Biology and Evolution 24 (10) (2007) 2266–2276.
[40] Casey W. Dunn, et al., Broad phylogenomic sampling improves resolution of the animal tree of life, Nature 452 (2008) 745–750.
[41] J.A. Tenreiro Machado, António C. Costa, Maria Dulce Quelhas, Fractional dynamics in DNA, Communications in Nonlinear Science and Numerical Simulations 16 (8) (2011) 2963–2969.
[42] Alan V. Oppenheim, Alan S. Willsky, S.Hamid Nawab, Signals and Systems, 2nd ed., Prentice Hall, 1996.