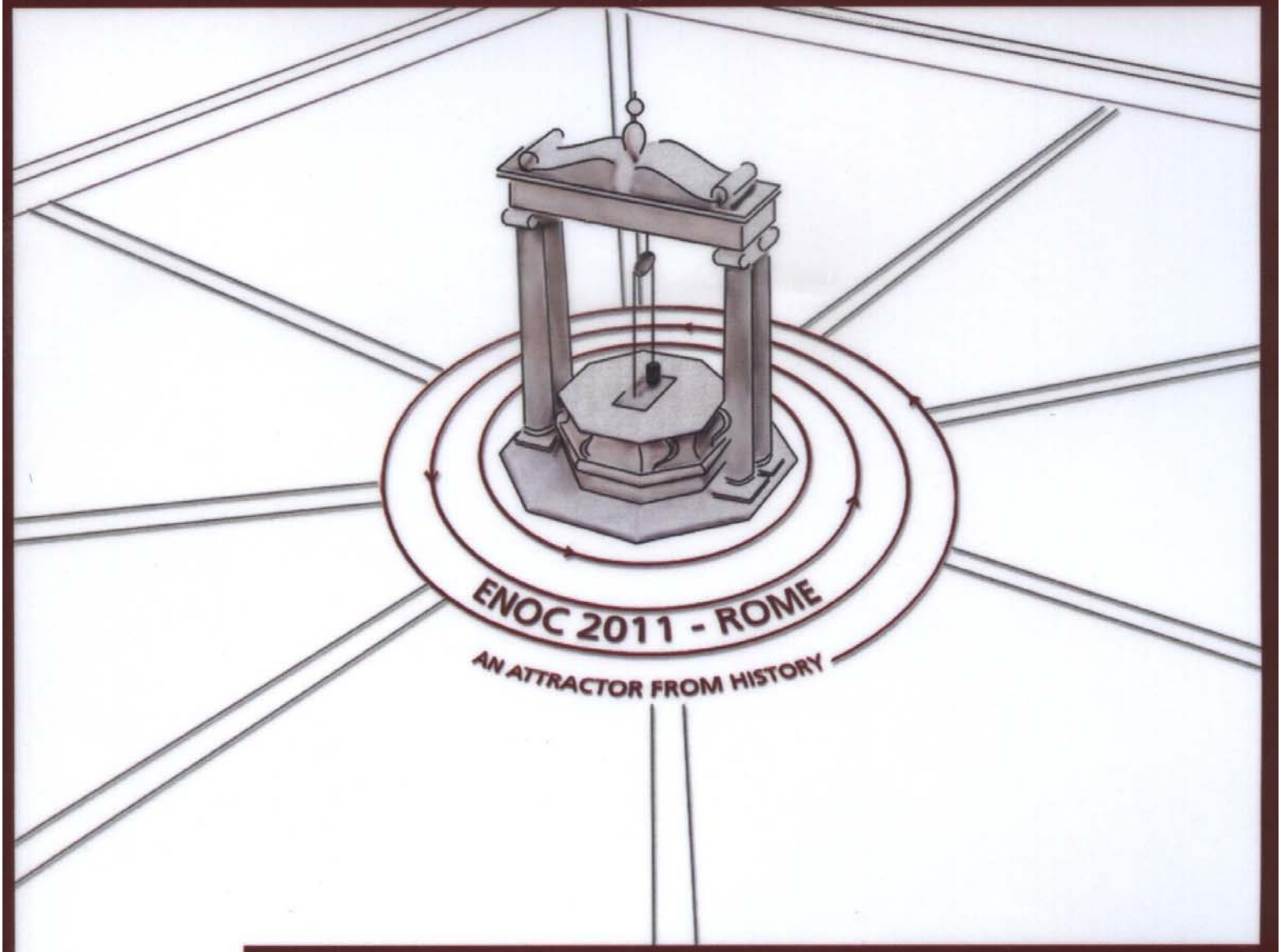

DIPARTIMENTO DI INGEGNERIA
STRUTTURALE E GEOTECNICA

SAPIENZA
UNIVERSITÀ DI ROMA

BOOKLET OF ABSTRACTS

Eds: D. Bernardini, G. Rega and F. Romeo

ENOC 2011

7th EUROPEAN NONLINEAR DYNAMICS CONFERENCE
July 24 - 29, 2011 Rome

DNA code analysis with fractional calculus

J. A. Tenreiro Machado*, António C. Costa** and Maria Dulce Quelhas ***
*Institute of Engineering of Porto, Dept. of Electrical Engineering, Porto Portugal
**Institute of Engineering of Porto, Dept. of Informatics Engineering, Porto Portugal
*** National Health Institute, Medical Genetics Center, Porto Portugal

Summary. This paper addresses the DNA code analysis in the perspective of dynamics and fractional calculus. Several mathematical tools are selected to establish a quantitative method without distorting the alphabet represented by the sequence of DNA bases. The association of Gray code, Fourier transform and Fractional calculus leads to a categorical representation of species and chromosomes.

Introduction

This paper studies the deoxyribonucleic acid (DNA) code in the perspective of system dynamics. A close observation of the DNA structure leads to the conclusion that "dynamic tools" may prove to be powerful allies in this endeavour. This observation motivated the association of logical and mathematical concepts namely, Gray coding, Fourier transform (FT) and fractional calculus (FC) for the analysis of the DNA data of twenty species. The results reveal important relationships between chromosomes (Chrs) and species, pointing to the goodness of the proposed methodology, and motivating further research with the usual formalisms of system dynamics. This paper presents briefly the mathematical tools and formulates their application in the framework of the DNA sequence decoding [1, 2, 3, 4, 5].

Mathematical Methods and DNA Decoding

DNA is made up of two polymers forming a double helix structure. The polymers contain four different nitrogenous bases: thymine, cytosine, adenine, and guanine, represented as T, C, A, and G. Each type of base on one strand forms a bond with just one type of base on the other strand, with A bonding only to T, and C bonding only to G. From the available DNA sequences a substantial part is organized into Chrs and has been used in this study. For converting the DNA code into a numerical value it is observed that we are handing an alphabet with symbols "T", "C", "A", "G". The available data includes a fifth symbol, represented by "N", which has no practical meaning for the DNA coding and, therefore, this symbol was considered as "zero" during the calculations. We have different values when considering DNA sequences with length ranging from $n = 1$, representing a counting of $m = 4^1$ states, up to $n = 5$, representing the dynamics of a system with $m = 4^5$ states. It must be noted that we are handling non-numerical quantities. Therefore, in order to prevent inserting a numerical order, it was decided to adopt numerical values according to the binary Gray encoding applied to the DNA alphabet. Therefore, we get the sequences $\{A\}$ $\{C\}$ $\{G\}$ $\{T\}$, and $\{AA\}$ $\{AC\}$ $\{AG\}$ $\{AT\}$ $\{CT\}$ $\{CG\}$ $\{CC\}$ $\{CA\}$ $\{GA\}$ $\{GC\}$ $\{GG\}$ $\{GT\}$ $\{TT\}$ $\{TG\}$ $\{TC\}$ $\{TA\}$ for $n = 1$ and $n = 2$, respectively. Furthermore, for the Gray code sequence conversion, a collection of windows was adopted with an overlapping of $n - 1$ consecutive bases. The numerical output of the DNA encoding is given by $y = sin(2x/m)$ where $x = 0, 1, ..., m - 1$ for the consecutive sequences of n symbols in the Gray encoding. In the paper it was decided to analyse eleven mammals, two birds, two fishes, two insects, two nematodes and one fungus, namely, Human (Hu), Common Chimpanzee (Ch), Orangutan (Or), Rhesus monkey (Rm), Pig (Pi), Opossum (Op), Mouse (Mm), Rat (Rn), Dog (Do), Cow (Co), Horse (Eq), Chicken (Ck), Zebra Finch (Tg), Zebrafish (Zf), Tetraodon (Tn), Gambiae mosquito (Ag), Honeybee (Am), Caenorhabditis elegans (Ce), Caenorhabditis briggsae (Cb), and Yeast (Sc).

DNA Analysis with the Fourier Transform

The combination of Gray encoding and trigonometric function was applied to the Chrs of the twenty species and, for each case, the FT was calculated. It was observed that the amplitude of FT versus the frequency ω could be approximated by a power function $a\omega^b$, $a \in \mathbb{R}, b \in \mathbb{R}^+$, with the parameters (a, b) to be determined by a least square fit procedure. The FTs were evaluated for all species with $n = \{1, 2, 3, 4, 5\}$ and the corresponding power law trendlines were obtained. In the locus of (a, b) was observed that, for each Chr, the corresponding trace moves from bottom to top and from left to right when varying from $n = 1$ up to $n = 3$, while for $n = \{4, 5\}$ it remains essentially in the same location. Due to this property in the sequel was considered solely the $n = 3$ encoding. Once clarified these aspects it was performed the FT calculation and the power law approximation for the Chrs of the twenty species. Figure 1 shows the locus of (a, b) parameters for the 415 chromosomes. The parameter a is related to the "energy" which reflects partially the size of the Chr. Therefore, we observe a tendency for smaller/larger values of the point labels in the right/left of the locus of (a, b). The parameter b is related with the "information content", being more close/apart to/from zero as the DNA is more random/correlated along the sequence. We verify that mammals have more negative values of b. We note also a separation both in the perspective of species and Chrs. In terms of species, we observe at the top left side a cluster constituted by the Sc, followed by the group Rm, Tn, Am, Ce, and Cb. Somewhat lower to the right we have the main part of the mammals namely the Hu, Ch, Or, Pi, Mm, Rn, Do, Co, and Eq. Somewhat peculiar is the place of Op separated to the right from the rest of the mammals. The birds Ck and Tg are in the middle of the two groups covering all range from left to right. The Zf and the Ag superimpose partially in the mammals. In terms of Chrs we observe particularly in mammals that, in general,

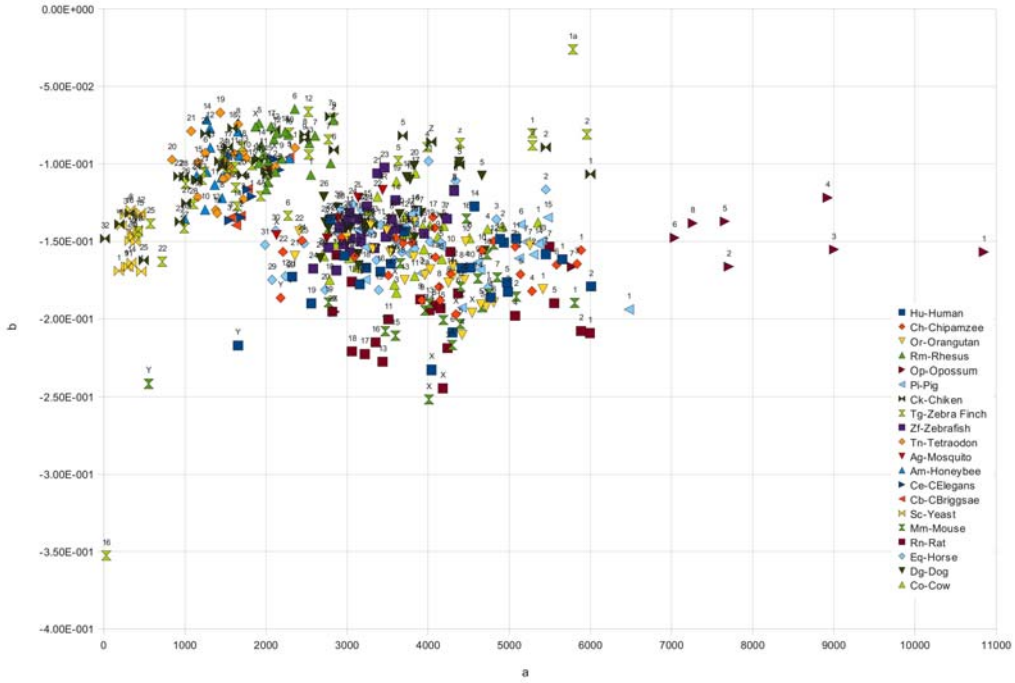

Figure 1: Locus of the parameters (a, b) for the 415 chromosomes of the twenty species when $n = 3$.

Chrs with the same numbering are relatively close. For example Chrs 1 and 3 are very similar for Human, Chimpanzee and Orangutan. Nevertheless, Chrs 2, X and Y (when it exists) reveal a remarkable difference. Obviously, much more can be extracted from the locus of (a, b) with 415 points and a more detailed analysis will be developed in the future.

Acknowledgements

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.cgi GENOME=Pongo%20abelii
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. 2005 Dec 8;438:803-19
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/
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 - Sacchromyces Genome Database, http://www.yeastgenome.org/

References

[1] Murphy W.J., Pringle T.H., Crider T.A., Springer M.S., Miller W. (2007) Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Research* **17**:413-421.

[2] Pearson H. (2006) Genetics: what is a gene? *Nature* **441**:398-401.

[3] Sims G.E., Jun S. , Wu G. A., Kim S. (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions *Proc. of the National Academy of Sciences of the United States of America* **106**:2677-2682.

[4] UCSC Genome Bioinformatics, web site http://hgdownload.cse.ucsc.edu/downloads.html

[5] Machado J.A., Costa A.C., Quelhas M.D. (2011) Fractional dynamics in DNA. *Communications in Nonlinear Science and Numerical Simulations* **16**:2963-2969.