ON THE DNA OF ELEVEN MAMMALS

J. TENREIRO MACHADO*
Dept. of Electrical Engineering, Institute of Engineering of Porto,
Rua Dr. António Bernardino de Almeida, 431
Porto, 4200-072, Portugal
jtm@isep.ipp.pt

ANTÓNIO C. COSTA
Dept. of Informatics Engineering, Institute of Engineering of Porto,
Rua Dr. António Bernardino de Almeida, 431
Porto, 4200-072, Portugal
acc@isep.ipp.pt

MARIA DULCE QUELHAS
Medical Genetics Center, National Health Institute,
Praça Pedro Nunes, 88
Porto, 4099-028, Portugal
mdquelhas@gmail.com

Received (to be inserted by publisher)

This paper studies the DNA code of eleven mammals in 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 non-integer 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 the modelling of a considerable number of phenomena characterized by long range memory properties [Hilfer, 2000], [Machado et al., 2011], [Magin, 2006], [Mainardi, 2010], [Monje et al., 2010], [Oustaloup, 1991], [Sabatier et al., 2007], [Zaslavsky, 2005]. In fact, FC emerged as the the key concept for the study of dynamical systems where classical tools reveals strong limitations. Due to this state of affairs we verify that presently the application of FC concepts encompass a wide spectrum of fields going from physics [Baleanu

*Dept. of Electrical Engineering, Institute of Engineering of Porto, Rua Dr. António Bernardino de Almeida, 431, Porto, 4200-072, Portugal.
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., 2004], [Emanuele II et al., 2005], [Jeng et al., 2006], [Sims et al., 2009], [Yin & Yau, 2005] in the perspective of fractional dynamics. It is believed that, for a given living being, besides an information level about its “structural construction”, DNA also includes other levels of information such as the history of evolution up to the present state, or instructions for the behaviour of each individual during its lifetime [Afreixo et al., 2004], [Dunn et al., 2008], [Leitão et al., 2005]. These distinct time scales reveal that we are in the presence of a complex code and that tools usual in 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, for the analysis of 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 analyses the relationship between chromosomes and species for a set of eleven mammals. Finally, section 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 twenty-three pairs of separate DNA-protein complexes (chromosomes), each containing, on average, 160 million nucleotide pairs. This massive amount of information is being collected during the last years, as the result of the 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 translation of the four symbols to a numerical values must be careful in order to prevent, from inception, any improper effect that will pervade the rest of the numerical treatment. In a previous paper [Machado et al., 2011] it was considered the adoption of Gray code and multiple length sequences. In this paper it is investigated a simpler process corresponding to having the direct symbol translation:

\[
A = 1 + i0, C = -1 + i0, T = 0 + i, G = 0 - i, N = 0 + i0
\]  

(1)

where \(i = \sqrt{-1}\).

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.

Having defined the translation scheme that generates numerical values, the second logical step to establish 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” \(x(t)\) being \(t\) denoted as the “time” in the absence of of a better term. We must observe that we are not referring to any value in units of seconds, but solely freely describing the consecutive base sequencing in the DNA code. The third phase of processing 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:

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

(2)

where \( F \) represents the Fourier operator and \( \omega \) will be named as “frequency”.

We decided to analyse eleven mammals (Ebersberger et al., 2007), (Murphy et al., 2007), (Prasad & Bourque, 20089), (Zhao & Bourque, 2009) namely, Human, common Chimpanzee, Orangutan, Rhesus monkey, Pig, Opossum, Mouse, Rat, Dog, Cow, Horse. The chromosomes characteristics of each DNA species are presented in Table 1. It should be noted that there is presently no theory, or even some empirical understanding, about the number of chromosomes or their length. The chromosome numbering follows simply a naive classification by size, being chromosome one the largest.

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

<table>
<thead>
<tr>
<th>Species</th>
<th>Number</th>
<th>Chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>24</td>
<td>1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, Y</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>25</td>
<td>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</td>
</tr>
<tr>
<td>Orangutan</td>
<td>24</td>
<td>1, 2a, 2b, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X</td>
</tr>
<tr>
<td>Rhesus</td>
<td>21</td>
<td>1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, X</td>
</tr>
<tr>
<td>Pig</td>
<td>19</td>
<td>1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, X</td>
</tr>
<tr>
<td>Opossum</td>
<td>9</td>
<td>1, 2, 3, 4, 5, 6, 7, 8, X</td>
</tr>
<tr>
<td>Mouse</td>
<td>21</td>
<td>1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, X, Y</td>
</tr>
<tr>
<td>Rat</td>
<td>21</td>
<td>1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, X</td>
</tr>
<tr>
<td>Dog</td>
<td>38</td>
<td>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</td>
</tr>
<tr>
<td>Cow</td>
<td>30</td>
<td>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</td>
</tr>
<tr>
<td>Horse</td>
<td>32</td>
<td>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</td>
</tr>
</tbody>
</table>

3. Fractional Phenomena

In section two it was established a set of eleven mammals making a total of 265 chromosomes. Therefore, according with the logical reasoning formulated previously, for each chromosome, a ‘signal’ \( x_i(t) \), \( i = 1, \ldots , 265 \), in the perspective of (1) was derived. Nevertheless, the application of the Fourier transform (2) leads to real and imaginary components depicting considerable perturbations. Therefore, it was adopted the amplitude of the Fourier transform, because it removes a considerable part of the ‘noise’.

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

\[
|F \{ x_i(t) \}| \approx a\omega^b.
\]

(3)
where $a > 0$ and $b$ are parameters to be determined by a least square fit procedure. For example, figure 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 calculate by averaging the values of the parameters $a$ and $b$. It is also represented the outer contour of the set of chromosomes for each species. The parameter $a$ reflects the size of the chromosome, while parameter $b$ describes that nature of the signal information. Therefore, we observe a tendency for smaller/larger values of the point labels in the right/left of the locus $(a,b)$. Furthermore, it is clear the fractional order of the values obtained for $b$.

![Fourier transform of the signal for the Human chromosomes X and Y and the corresponding power law approximation.](image)

**Fig. 1.** Fourier transform of the signal for the Human chromosomes X and Y and the corresponding power law approximation.
In order to compare the different species figures 13 to 14 depict the superposition of the contours and the centers of the polygons of the eleven mammals, respectively.

We verify that the Opossum is very different from the rest of the mammals. Moreover, we observe that the Chimpanzee and 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 Horse reveal strong similarities between themselves. Therefore, most of these results follow what is known qualitatively in known, but
Fig. 4. Locus of the parameters \((a, b)\) for the 24 chromosomes of the Orangutan.

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

provide a quantitative tool for measuring ‘distances’ between species in the viewpoint of phylogenetics. In a different level, the results demonstrate that the information ceded in the DNA has long memory and that characteristic is easily captured with the tools of FC.
4. Conclusions

In this paper the chromosomes four symbol alphabet was converted to a numerical value 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
Fig. 8. Locus of the parameters \((a, b)\) for the 21 chromosomes of the Mouse.

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

new research efforts in the perspective of FC.

Acknowledgements

The authors thank the following organizations for allowing access to genome data:
Fig. 10. Locus of the parameters \((a, b)\) for the 38 chromosomes of the Dog.

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

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

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

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

http://www.broad.mit.edu/mammals/dog/,

- Horse - The Broad Institute, http://www.broad.mit.edu/mammals/horse/
- Opossum - The Broad Institute, http://www.broad.mit.edu/mammals/opossum/
- UCSC Genome Bioinformatics http://hgdownload.cse.ucsc.edu/downloads.html

References

REFERENCES


profiles (FFP) and optimal resolutions,” *Proc. of the National Academy of Sciences of the United States of America* 106, 2677–2682.


