Investigation of the fractality and mutual information of DNA sequences: The case of human B-myosin

Claudia Sierra*
Microbiology Lab, Faculty of Science, UNAM, New Mexico, Mexica

Murat Tuğrul†
Institute of Cross-Disciplinary Physics and Complex Systems (CSIC-UIB), E-07122 Palma de Mallorca, Spain

(Dated: August 9, 2008)

In general this project serve the authors as an initiative study for "informational theoretic approach to biological sequences". First of all, we went through Wavelet Transform Maxima Method for investigating the fractality of a signal and apply it to a DNA walk yielded from gene B-myosin in Human. Secondly, we held the basics of the information theory and in special Mutual Information.

PACS numbers:

INTRODUCTION

One can try to approach to biological evolution by means of informational theoretic [1]. Such an attempt obviously needs a well background in information theory and a good unerstading of biological sequences. Accordingly, this project work will be simply a basic work for the authors to initiate studying in corelated disciplines which is especially held in the summer school.

It might be worthwhile here to state some questions at the begining of the project in order give a sight to reader about the content of the project,

- How can we measure the information content embeded into DNA sequences. (Shannon Entropy, Chaitin algorithmic information, mutual information, etc.)
- Is there a long range probabibilistic dependence in the DNA sequence and if so how can we overcome this computational problem when we measure the information content?
- What are the informational features differences between randomly generated (reshuffling, or under a dynamical mechanism) and real DNA sequences?

METHODS & DATA

At the following subsections you will find the basic theories and used data.

Wavelet Transform Maxima Method (WTMM)

The use of Wavelet Trasnform (WT) for detecting the (multi-)fractality was well summarized by Arneodo et al.Ref. [2] as "The WT has been early recognized as a mathematical microscope that is well adopted to reveal the hierarchy that governs the spatial distribution of singularities of multifractal measures." Using the notation of the same reference, WTMM can be explained as follows,

Wavelet Transform:

$$T\psi[f](x_o, a) = a^{-1}\int_{-\infty}^{\infty} f(x)\psi(\frac{x-x_o}{a})dx \qquad (1)$$

where x_o and a are respectively space and scale parameter. ψ function (in this context) is taken such that it gives zero mean and obeys the orthaganility conditions with polynomial functions of order m, i.e.,

$$\int_{-\infty}^{\infty} x^m \psi(x)dx = 0, 0 \le m \le n_\psi \qquad (2)$$

So in general applications people commonly use the derivatives of the gaussian function,

$$g^{(N)}(x) = \frac{d^N}{dx^N} e^{-x^2/2} \qquad (3)$$

$$\begin{aligned} T_{g^{(N)}}[f](x,a) &= a^{-1}\int_{-\infty}^{\infty} f(y) g^{(N)}((y-x)/a)dy \qquad (4) \\ &= a^N \frac{d^N}{dx^N} T_{g(0)}[f](x,a) \qquad (5) \end{aligned}$$

Remembering $n_\psi = N$, $T_\psi[f](x,a) \sim a^{h(x_o)}$ as a goes to 0 if $n_\psi \ge h(x_o)$. The main idea arised here as determining the Holder exponent $h(x_o)$ as the slope of log-log plot of the WT amplitude versus the scale parameter a. But for the simple calculations Arneodo et al. used an anology to statistical physics with defining a partition function,

$$Z(q,a) = \sum_l (sup_{(x,a') \in l} |T\psi[f](x,a')|)^q \qquad (6)$$

where l is a maxima lines in WT, $q\epsilon\Re$, $a' < a$ and sup is a way of defining “Hausdorff-like” partition which yileds

$$Z(q,a) \sim a^{\tau(q)}, a \to 0^{+} \tag{7}$$

One can pass from here to the singularity spectrum by using Legendre transform, i.e.

$$D(h) = min_{q}(qh - \tau(q)) \tag{8}$$

Information Theory & Mutual Information

The amount of the information can be quantified for example by using Shannon Entropy, i.e.

$$H = -\sum p(i)log(p(i)) \tag{9}$$

however, just a single number might be meaningles in many context. On the other hand, a quantity which measure the mutual amount of information can be more meaningful. In other words, mutual information would give us a information notion between object and its surronding which is expected to be more valuable in biological systems.

$$MI = -\sum p(i,j)log(\frac{p(i)p(j)}{p(i,j)}) \tag{10}$$

where p(i), p(j) and p(i,j) stand for probabilities of i and j and joint probability [3].

After some algebra one can show

$$MI = H(p(j)) - < H(p(j|i)) >_{i} \tag{11}$$

where $p(j|i)$ is the conditional probability with given i. (Remember $p(i)p(j|i) = p(i,j)$)

a simple model for applying Mutual Information

Let us consider two sequence ensembles A and B where i and j represents the sequence index respectively. We will consider the p(j) time independent (let us say for this project according to statistics of gene β Myosin). We will start p(j—i) as random accordingly p(i) and consider them time dependent under the dynamics. More realistic dynamical rules can be driven into model but at the moment let us rule that at every discrete time step t, the interactions between A and B are mutated and so p(j—i) and if these mutations has tendency to increase the MI, the dynamics will accept the mutations and otherwise reject them.

Data

For the sake of investigation one real gene we use the exon and intron of β-Myosin in Human by obtaining it from the website $http://www.ncbi.nlm.nih.gov/$. It has 6044 nucleotides.

RESULTS

First of all, in order to investigate the fractality of the gene sequence, we created a DNA walk by summing distance from the origin by adding +1 to f_n when we see A or C at n^{th} position and subtracting -1 viceversa. Generated 1-D DNA walk is presented in Fig. 1.

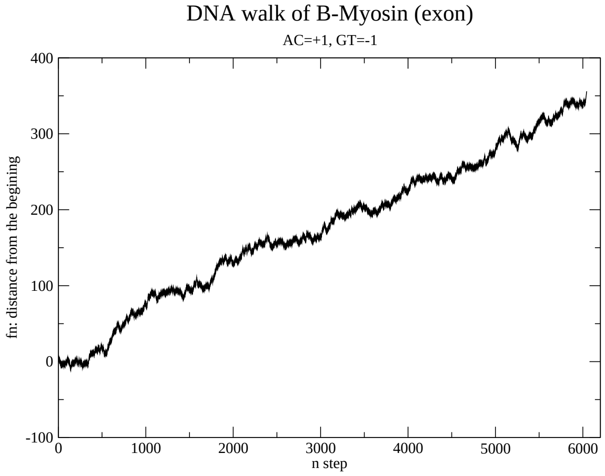

FIG. 1: DNA walk of β-Myosin

We analyzed its WT and WTMM by using Matlab and Fraclab and show the results in Fig. 2 and Fig. 3, respectively. According to results, DNA walk held by the exon part of the β-myosin shows the multifractality where the spectrum is shown at rigth botom of Fig. 3. However, one should be aware of the method used here might be very dependent of the parameters and for more concrete results one should repeat the procedure by using other softwares available such as LastWave [4] provided by Arneodo and his collegues.

Secondly, We we investigated basic probabilities in the gene β-myosin. We saw p(G), p(C), p(A) and p(T) are 0.311714, 0.260093, 0.269358 and 0.158835, respectively. Using these probabilities one can estimate two letter probabilities and compare it with real case. For example, we expect from a random sequence to hold p(AC)=p(A)p(C)= 0.070058.

Thirdly, We planned to investigate the change of the MI of the system proposed in the method section, i.e. we ask $\frac{dMI}{dt}$. The reason of such investigation is in biological

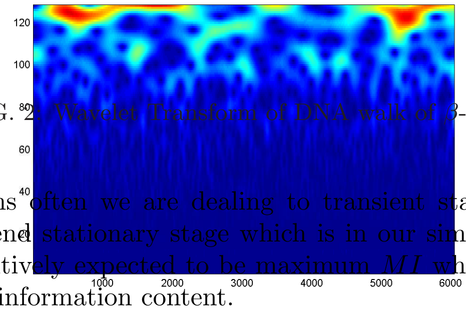

FIG. 2: Wavelet Transform of DNA walk of β-Myosin

systems often we are dealing to transient stage rather than end stationary stage which is in our simple model is intutively expected to be maximum MI which means equal information content.

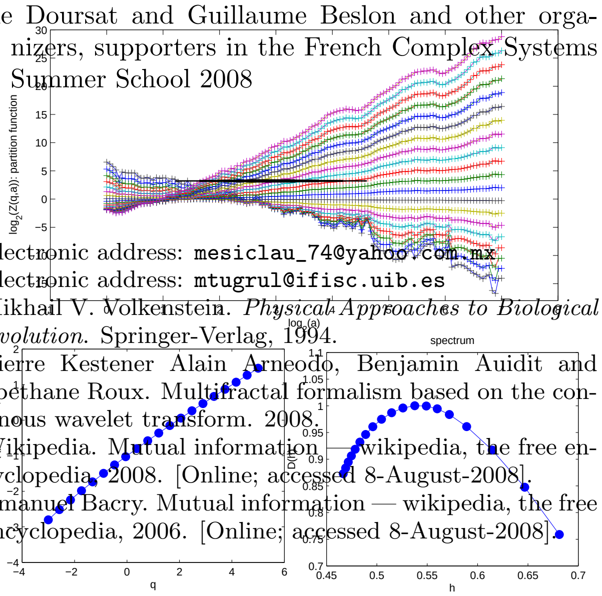

FIG. 3: WTMM of DNA walk of β-Myosin

$$\frac{dMI}{dt} = \frac{d}{dt}(H(p(j)) - < H(p(j|i)) >_i) \qquad (12)$$

$$= - < \frac{d}{dt}(H(p(j|i))) >_i \qquad (13)$$

DISCUSSION

We obtained that DNA walk of β-myosin shows a multifractal structure as shown in Fig. 3. Biological reading of this result is that the gene we hold includes long range correlation. But for more concrete conclusion about this result might need more study on the topic.

ACKNOWLEDGEMENT

MT and CS cordially thank

supervisors of this project: Pierre Baudot, Nicolas Brodu and Peter Grassberger,

all participants and lecturers,

Rene Doursat and Guillaume Beslon and other organizers, supporters in the French Complex Systems Summer School 2008

[*] Electronic address: mesiclau_74@yahoo.com.mx

[†] Electronic address: mtugrul@ifisc.uib.es

[1] Mikhail V. Volkenstein. *Physical Approaches to Biological Evolution*. Springer-Verlag, 1994.

[2] Pierre Kestener Alain Arneodo, Benjamin Auidit and Spethane Roux. Multifractal formalism based on the continous wavelet transform. 2008.

[3] Wikipedia. Mutual information — wikipedia, the free encyclopedia, 2008. [Online; accessed 8-August-2008].

[4] Emanuel Bacry. Mutual information — wikipedia, the free encyclopedia, 2006. [Online; accessed 8-August-2008].