Poincaré Recurrences and Multifractal Properties of Genomic Sequences

L. Rossi, G. Turchetti

University of Bologna, Department of Physics, Bologna 40126, Italy

Abstract

We propose the return times spectra as a tool to analyze the genomic sequences. The spectra discriminate between regular, chaotic and mixed dynamical systems since generally their decay follows respectively an exponential law, a power law or is a linear combination of them. The analysis of words of $n$ bases from a DNA sequence exhibits the same exponential behavior for the coding and noncoding component. The multifractal analysis shows a definite difference, which suggests a less uniform structure for the noncoding part.

Key words: Genomic sequences, Poincaré recurrences, Multifractal analysis

1 Introduction

The aim of our work consists in the analysis of the coding and noncoding regions of genomic sequences by the statistics of Poincaré recurrences, a technique which, in the recent years, has become increasingly relevant in the theory of dynamical systems.

Although many studies has been performed on genomic sequences using statistical methods (see for example [1-5]), the use of Poincaré recurrences is new. This choice was made because it seems capable to capture some fundamental features of the dynamics of the underlying systems. For instance, it is known that for a wide class of systems having a strongly mixing dynamics [6] the spectrum of return times decays exponentially; on the other hand, recently it has been shown [7] that for a particular integrable system the distribution of Poincaré recurrences follows an algebraic decay law.

Email addresses: lux.ruber@libero.it (L. Rossi), turchetti@bo.infn.it (G. Turchetti).

Preprint submitted to Elsevier Science 22 November 2004
Furthermore, [7], it has been proved that for a class of mixed systems (i.e., systems composed of two or more invariant regions with respect to the dynamics) the spectrum is a linear combination of the spectra characteristic of each invariant region.

2 Genomic sequences

The genomic sequences that we used in our analysis were chosen from the set of complete genomes located at the GenBank database (www.ncbi.nlm.nih.gov). We considered both eukaryotic (about 60 chromosomes) and bacterial sequences: among the organisms we studied there are *C. Elegans, D. Melanogaster, S. Cerevisiae* and *E. Coli*.

The coding sequences were obtained as follows: from a given genomic sequence (usually a chromosome) we extracted all the coding regions and then glued them together, in the same order in which they appeared in the original sequence; the noncoding sequences were constructed in a similar way.

3 Some statistical analysis

In the beginning, we performed a series of analysis to understand some important statistical properties of the coding and noncoding sequences. Considering them as made of contiguous, non overlapping groups of $n$ adjacent nucleotides (such groups will be called “words”) we computed the distribution (spectrum) of the frequency of occurrence of words of a given length $n$, with $n$ between 1 and 8.

The spectra obtained for pairs of nucleotides ($n = 2$) exhibited the presence of correlations among adjacent nucleotides in both kind of sequences, whereas the spectra concerning words of larger length showed an interesting difference between the coding and noncoding part of the genomic sequences considered. Indeed in the noncoding case there is generally a quite small group of words having a significantly higher frequency of occurrence with respect to all other possible words, while in the coding case the distribution appears to be more uniform.

This fact is in agreement with what was obtained by analyzing the multifractal properties of the measure associated to the sequences, such a measure being constructed through the distribution of the frequency of occurrence of the words. Since the multifractal spectrum $f(\alpha)$ of the coding regions is usually narrower than that of the noncoding ones (above all for eukaryotic organisms),
Fig. 1. Multifractal spectra of the coding and noncoding sequences from the chromosome III of *C. Elegans*, obtained for words made by seven nucleotides.

this means that the coding part is composed by words distributed more uniformly. An example of such a different behavior is shown in Fig. 1, in which we report the multifractal spectra related to the coding and noncoding regions obtained from the chromosome III of *C. Elegans*.

Another interesting result came from the computation of the Shannon entropy (which is proportional to the degree of compressibility of a string of symbols), because it generally was nearly the same for both kind of sequences.

### 4 Poincaré recurrences

In order to study the genomic sequences through the statistics of Poincaré recurrences, it is essential to construct from each of them a dynamical system. This involves the definition of a so called phase space $\mathcal{X}$, that is the set of the states of the system, of a dynamics $T$, which describes how a state is transformed into another, and of a measure $\mu$.

We choose as our phase space the collection of all the possible words of a given fixed length. The dynamics $T$ is defined in such a way it associates to a word $x$ the word which immediately follows the first occurrence of $x$ in the sequence; if $x$ is not present, it simply is ignored. Of course other definitions for the dynamics could equally be considered, for example by associating to a word $x$ the word obtained from a weighting process concerning *all* the occurrences of $x$ in the sequence. The measure $\mu$ of a subset of $\mathcal{X}$ is given by the ratio
between the number of words belonging to that subset and those of the whole phase space.

The return time $\tau_A(x)$ into a domain $A \subseteq \mathcal{X}$ of a word $x \in A$ is defined as follows

$$
\tau_A(x) = \min \left( \left\{ k \in \mathbb{N} : T^k(x) \in A \right\} \cup \{+\infty\} \right),
$$

while the mean return time $\langle \tau_A \rangle$ into $A$ is given by

$$
\langle \tau_A \rangle = \int_A \tau_A(x) \, d\mu_A,
$$

where $\mu_A$ denotes the conditional measure on $A$.

Through the previous definitions we can now introduce the notion of return times spectrum

$$
F_A(t) = \mu_A \left( \left\{ x \in A : \frac{\tau_A(x)}{\langle \tau_A \rangle} > t \right\} \right).
$$

Roughly speaking, the return times spectrum $F_A(t)$ represents the fraction of the words of $A$ which still did not return into $A$ after the time $t \langle \tau_A \rangle$.

The main result we obtained is that, regardless of the word’s length chosen for the computation of the statistics of Poincaré recurrences, the spectrum essentially is the same for both kind of sequences. Furthermore the spectrum seems to follow, with a very good accuracy, an exponential decay law (see Figg. 2, 3) as if the genomic sequences were composed by completely uncorrelated nucleotides.

From the point of view of the Poincaré recurrences the coding and noncoding regions behave thus like strongly mixing dynamical systems, which are characterized by a rapid decay of the correlations. It seems therefore sensible to conclude that the weight of the long-range correlations is scarce, at least compared to the short-range correlations.

5 Acknowledgments

This work was supported by a grant of the Italian Ministry for the University and the Scientific Research entitled A multidisciplinary approach to the regulation of the immune system.
Fig. 2. Poincaré recurrences spectra obtained from the coding sequences of some organisms, for words seven nucleotides long; the dotted line represents the function $\exp(-t)$.

Fig. 3. Poincaré recurrences spectra obtained from the noncoding sequences of some organisms, for words seven nucleotides long; the dotted line represents the function $\exp(-t)$.

References


