Research

Author for correspondence:
S. Srivastava
e-mail: shambhaviuoa@gmail.com

THE ROYAL SOCIETY
PUBLISHING

Markovian language model of the DNA and its information content

S. Srivastava and M. S. Baptista

Institute for Complex Systems and Mathematical Biology, SUPA, University of Aberdeen, Aberdeen AB24 3UE, UK

This work proposes a Markovian memoryless model for the DNA that simplifies enormously the complexity of it. We encode nucleotide sequences into symbolic sequences, called words, from which we establish meaningful length of words and groups of words that share symbolic similarities. Interpreting a node to represent a group of similar words and edges to represent their functional connectivity allows us to construct a network of the grammatical rules governing the appearance of groups of words in the DNA. Our model allows us to predict the transition between groups of words in the DNA with unprecedented accuracy, and to easily calculate many informational quantities to better characterize the DNA. In addition, we reduce the DNA of known bacteria to a network of only tens of nodes, show how our model can be used to detect similar (or dissimilar) genes in different organisms, and which sequences of symbols are responsible for most of the information content of the DNA. Therefore, the DNA can indeed be treated as a language, a Markovian language, where a 'word' is an element of a group, and its grammar represents the rules behind the probability of transitions between any two groups.

1. Introduction

One of the most studied complex systems in biology is the genome of living organisms, composed of deoxyribonucleic acid (DNA). The central dogma of life is related to how DNA transcribes into mRNA which finally translates into proteins. A big challenge is to understand the complexity of the dynamical organization of the DNA, a system that is the result of millions of years of evolution. The genome of any organism is its hereditary information. It is encoded either in the form of DNA composed of four different types of chemical molecules (adenine (A), guanine (G), cytosine (C) and thymine (T)) or in the form of ribonucleic acid (RNA) as present

in many viruses. The genome includes both the genes and the non-coding sequences of the DNA/RNA [1]. Sequencing and high-throughput experiments have contributed much to the genomic data in the last 10–15 years. Still the data have to be analysed [2].

Natural language analysis has been a topic of interest in the last decade [3–5]. Natural language written texts can be considered as being composed of a series of letters, syllables, words or phrases. During the nineteenth century, many linguists like Schleicher and Haeckel interpreted language as a living system [6]. Based on this concept, Darwin also proposed that evolution of species and language are similar [7]. Many researchers have introduced the concept of linguistics into biology [8]. Brendel *et al.* [9] used formal linguistic concepts to define a basic grammar for genes, based on the idea that mutating a piece of genetic information was similar to modifying words. Similar to works that aimed at finding the relevant words, their relationships and their information content in natural languages, many studies have focused on analysing genomic sequences like DNA and proteins as if they were a language, using similar methodological approaches as the one used to model natural languages [8]. Formal linguistic concepts were used by Brendel *et al.* [9] to define basic grammatical rules that describe how genes can mutate, inspired in the grammatical rules of languages that regulate how and which word follows a previous word. Gramatikoff *et al.* [10] have used lexical statistics to identify and represent structural, functional and evolutionary relationships for multiple genomic sequences and texts from natural languages.

DNA sequences have also been analysed using approaches to characterize complex systems, for example, by converting the DNA sequences to numerical signals using different mappings [11–14]. A commonly used mapping is to convert the DNA into binary sequences [15]. Other mappings of the DNA look for spatial patterns by considering inter-nucleotide distances [11,16,17]. These models of the DNA capture the recurrence property of codons (a word formed by three nucleotides) [17] by measuring the statistics of the symbolic distance separation between two codons. In the work of Hao & Xie [18], they have shown how to analyse the long DNA sequences by converting it to an image. The DNA was characterized by the fractal-like patterns appearing in this image as a result of the forbidden words.

Our model is constructed using some concepts and tools from Ergodic Theory and Information Theory to interpret genomic data (nucleotide sequences) as if it were a language that can be analysed by the tools of symbolic dynamics. Each language has its rules. Our motivation in this work is to propose and study a meaningful language for the DNA. To establish a language for the DNA, we specify the length of words and a set of relevant groups of words, and create a network of words from functional connections, linking how topological complexity in the functional networks arises and how this complexity is connected to the complexity of life (production of proteins). In order to achieve this goal, we analysed the genome of *Escherichia coli*, *Shigella dysenteriae*, *Rhodococcus fascians* and *Saccharomyces cerevisiae*. These organisms are commonly used model systems: their genome and genes are well known, well studied, and can be used to test mathematical approaches towards modelling the DNA. First, we represent the genome of these organisms on a symbolic space. The words of the genome are encoded in such a way that the nucleotide sequences are represented by real numbers that can be plotted in a symbolic space. This space allows a straightforward characterization of the DNA through informational quantities (Shannon entropy rate, mutual information rate (MRI) and statistical measures), and ergodic quantities (correlation decay, and transition probabilities). To group the words and specify their lengths, we find a partition of the symbolic space composed by N^2 equal boxes. A box is a region in this symbolic space whose points within encode and define a group of words of length $2L$ that are all formed by the same small sequence of symbols with length $2L_n$, where $L_n = \frac{1}{2} \log_2(N)$, $N = 4^i$, $i \in N$, and $L_n < L$. To create a grammatical network of words of the DNA, we set the nodes to represent groups of words and the edges to represent the grammatical rules governing the transition probabilities between group of words.

Our Markovian model of the DNA is constructed by searching for the existence of a Markov partition (see §3b and the electronic supplementary material) in the symbolic space for which the correlations of the points within boxes of this partition separated 'in time' decays to approximatively zero, where 'time' denotes the nucleotide distance ($2L_n$) between two words encoded by two points in distinct boxes. This finding allows us to see the DNA as an approximate Markov process in which the behaviour of the whole system can be approximately described by the transitions among the group of words of finite length. It was shown however by Arnedo *et al.* [19] that long words composed of more than 100 bp in a DNA separated by many nucleotides have slow power-law correlation decay. Complementing the results in [19], Buldyrev *et al.* [13] have analysed the DNA sequences (words) and have shown that there is no correlation between sequences of nucleotides with lengths between 10 and 100 bp. Our Markov model for the DNA has memoryless properties. It is however based in the relationship between groups of words, not between words as in [13,19]. The correct choice for grouping the words provide the memoryless

property of our model. The transformation of a system that has correlation into a memoryless symbolic system is a normal procedure to study chaotic systems by using a symbolic representation of its trajectory. The correlation between time separated points in chaotic systems or correlation between symbols created by a non-Markovian partition of the phase space decays to zero certainly after an infinitely long time [20, 21]. However, the symbolic sequences generated by Markov partitions of chaotic systems have sequences whose correlations decay to zero even for a finite and small time separation, and for words of finite and small length.

Then, along the lines of the work in [22], we apply our model to create a network representation of the DNA, where the nodes represent the different group of words and edges represent the probability of transition between two groups of words that are strongly correlated and that are separated by one nucleotide. Preserving the connections responsible for most of the information of the DNA, we create reduced network models of the DNA, which capture the most relevant features of it, and are able to identify hub groups of words, which have a similar function to the core words identified by Gerlach & Altmann [23] in natural languages. This network model of the DNA allows us to create a similarity measure to identify genes in various genomes. We analysed genomes of *E. coli*, *Sh. dysenteriae*, *R. fascians* and *Sa. cerevisiae*. We observed that *E. coli* and *Sh. dysenteriae* are very similar to each other and there were remarkable differences between *E. coli* and *Sa. cerevisiae*.

In our Markov model, we use finite-length words which have all the same length, in contrast to the work [24] that considers variable word length. Our model allows for an easy calculation of entropy rates from groups of words with short length, a quantity that is usually calculated over very long words demanding high computational costs. In our model of DNA, there are forbidden words as in [18] but no forbidden group of words, all words contain and are formed by the same small symbolic sequence length of $\log_2(N)$, and any two words separated by $\log_2(N)$ nucleotides are roughly decorrelated. There are N^2 groups of words.

This paper is organized as follows. Our Markov model of the DNA is presented in §3. In §3.1, we show how to define a word, and in §3.2, we show how to group them. The network description of the DNA with a comprehensive analysis of the words that are mostly correlated is presented in §4, and the validation of the model and a discussion about the relationship between words in our model and the biological function behind them is presented in §5.

2. Symbolic representation of the DNA and a partition

Symbolic dynamics is a way to represent an orbit of a dynamical system whose points belong to the real set by a sequence of symbols taken from a finite set. Assume that a dynamical system is represented by (X, T), where $X \in \mathbb{R}^d$ is a set and T is a transformation that operates on X. T maps X to itself. The set X is the set of all possible states of a system and the transformation T evolves the state of the system, X, in time [21,25]. The first step into encoding an orbit by a symbolic sequence is the specification of a partition. This can be done by dividing X into cells, where a point within a cell would represent a symbolic sequence. The main motivation for using symbolic dynamics comes from the idea of observing an infinite resolution orbit by making finite resolution observations (represented by a finite alphabet of symbols) and being able to describe the properties of the system being observed [26]. This symbolic sequence is then encoded into a sequence of numbers, such that a symbolic space can be constructed. In this space, the transformation representing how points are mapped preserves the main features of the transformation T of the original space.

For our problem of modelling the DNA, the symbols are given. Our main interest is to define a biologically relevant encoding that creates a symbolic space whose distance between points is a measure of similarity between two sequences of $2L$ symbols, here denoted as words. In order to have a two-dimensional symbolic space, we construct two symbolic sequences, called the *past* and the *future* symbolic sequences. We first represent the nucleotides by natural numbers: $(A, T, G, C) = (0, 1, 2, 3)$. Given a symbolic sequence in the DNA with a length of $2L$ and assuming $L = 3$, if the DNA sequence is AATCGT, then the past sequence is AAT and the future sequence is CGT. The integer representation of this DNA sequence is given by $(s_0, s_1, s_2.s_3, s_4, s_5) = (001321)$. The encoding into real numbers of a symbolic sequence is made as in the following, assuming we are at the position jth of the DNA (the location of the jth nucleotide) then, the past symbolic sequence is encoded by

$$\delta_j = \sum_{i=1}^{L} \frac{s_{j-i} 4^{L-i}}{4^L - 1},$$

(2.1)
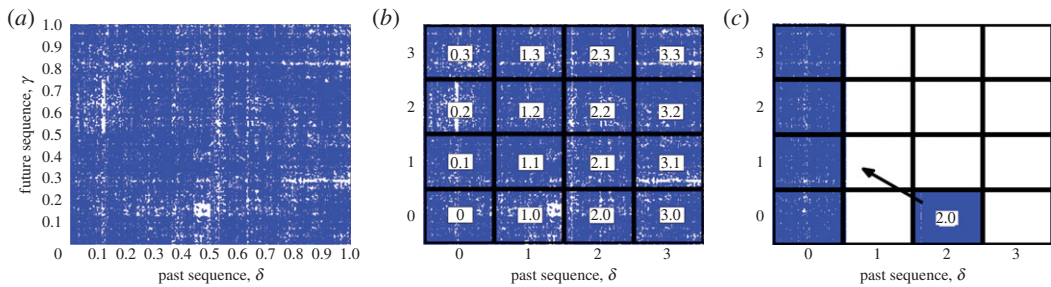
Figure 1. Symbolic space of the DNA of *E. coli* with horizontal axes representing the encoding of past sequences δ and vertical axes representing the encoding of future sequences γ. The grid lines represent a partition with $N = 4$. (*a*) Symbolic space of the DNA of *E. coli* where existing words are represented by blue points and white space represents the presence of forbidden words. (*b*) Symbolic space with a grid partition where the boxes define regions of points that encode similar words forming a group. (*c*) Symbolic space representing how points from a box move to four other boxes.

and the future sequence is encoded by

$$\gamma_j = \sum_{i=1}^{L} \frac{s_{j+i-1} 4^{L-i}}{4^L - 1}, \tag{2.2}$$

where 4 is the number of symbols encoding each nucleotide and $j \geq L$. A $2L$ length word centred at the position j is encoded by the point (δ_j, γ_j). This word is τ nucleotides apart from a word centred at $j + \tau$ that is encoded by the point $(\delta_{j+\tau}, \gamma_{j+\tau})$. We adopt a dynamic description for our symbolic space, meaning that we assume that the point (δ_j, γ_j) is mapped to the point $(\delta_{j+\tau}, \gamma_{j+\tau})$ after τ 'time' iterations. Both δ and $\gamma \in \text{IR} \, [0, 1]$. The encoding proposed in equations (2.1) and (2.2) is a standard encoding of a quaternary alphabet into a set of real numbers. The encoding of symbolic sequences into real numbers needs to satisfy some conditions in order to capture the underlying deterministic behaviour of the DNA: the symbolic space is independent on the starting point (so we can study equivalent spaces for the genes and the whole DNA or genome); the encoding produces similar results for palindromic (mirror images), which means that correlation decays as you go further into the past or further into the future; the encoding has finite-length words, creating a symbolic space whose distance between points in one coordinate is no smaller than 4^{-L} (or not smaller than $2^{-L}\sqrt{2}$ in the two-dimensional space); one-to-one coding, such that a symbolic sequence will be uniquely encoded into a real number.

A two-dimensional picture (figure 1*a*) showing all the point coordinates (δ, γ) is called *symbolic space of the DNA*. Figure 1 horizontal axes correspond to the encoding of the past sequences, δ, and the vertical axes correspond to the encoding of the future sequences, γ. We note some very dense areas in this space which means that many similar nucleotides sequences exist in the genome. There are some blank spaces, which represent forbidden sequences in the DNA. In figure 1*b*, we place a grid of N^2 boxes ($N = 4$) into the symbolic space. There are N columns and N rows. A column in this space represents boxes where points encode symbolic sequences that contain the same length L_n past symbolic sequence. A row represents all the boxes whose points encode symbolic sequences that contain the same length L_n future symbolic sequence. Each box contains points that encode a group of words that are similar, i.e. they all have the same length L_n past and future symbolic sequence. Let us understand the properties of symbolic trajectories in the symbolic space. In the partition of figure 1*b*, a point within the box with coordinates $\delta \in [0, 0.25]$ and $\gamma \in [0, 0.25]$ belongs to the box named '0.0'. Points within are mapped after 1 iterations (or 1 shift in the symbolic sequence) to the partition '0.X', where $X \in [0, 1, 2, 3]$. After two iterations they can be mapped everywhere in the symbolic space. Note that a point inside box '2.0' represents a symbolic sequence of length $2L$ in the DNA ($L_n < L$), where the first past symbol is '2' and the first future symbol is '0'. This sequence is centred at a location i of the DNA, has at the location $i - 1$ a nucleotide 'G' and at location i the nucleotide 'A'. Every $2L$ length symbolic sequence encoded by a point that belongs to the box '2.0' belongs to the group of words '2.0', having a $L_n = 1$ past symbol '2' and a $L_n = 1$ future symbol '0'. As illustrated in figure 1*c*, points from box '2.0' iterate to '0.X', where 'X' can be either 0, 1, 2 or 3. The length of the symbolic name of a box in a partition is given by $2L_n$, where $L_n < L$.

3. Determination of the approximate Markov partition: specification of word length and grouping

3.1. Determination of word length

The symbolic space allow us to find the hidden patterns of the DNA, since two similar symbolic sequences should be encoded by two nearby points in the symbolic space. We initially consider the DNA of *E. coli*. To have an estimation of the length of the words ($2L$) that our model should consider, we note that the genome size of *E. coli* is 4.6 million bp which means that it has 4 600 000 symbols. According to the encoding rules of equations (2.1) and (2.2), the maximum number of different sequences of length L would be $4\,600\,000 - 2L$. But for a given L, the total number of combinations possible is 4^{2L} sequences. Therefore, 4^{2L} should be at least smaller than 4 600 000 if we were to have each sequence to appear once. Then, $4^{2L} < 46\,00\,000$ and therefore $2L < 11.06$. Our main interest is to search for invariant properties in the language of the DNA, in order to obtain a stationary Markov model of it [27]. In order to achieve that, we determine the appropriate value of $L = L^*$ by the largest length that allows the topological entropy rate of the symbolic space to remain invariant as L is changed. If the length of the sequence is larger than L^* then the symbolic space properties change abruptly. Its statistical properties would no longer be invariant by a change in the length. As shown in the section 'Determination of word length' in the electronic supplementary material, the appropriate L^* to be used in this work is $L = L^* = 4$.

In contrast to the work of [24], we consider finite words with the same length, which allows for an easy calculation of entropy rates from groups of words of small short length, a quantity that is usually calculated over very long words demanding high computational costs.

3.2. Determination of grouping of words and the order of partition

A group is a set of symbolic sequences whose encoding (δ, γ) points fall within the same box of the partition. Let us now determine the optimal partition. In the symbolic spaces of the DNA, the minimum distance of points is equal to 4^{-L}, i.e. the minimum distance between points in a horizontal or vertical direction. We partition the symbolic space in N^2 equal boxes with sides of $\epsilon = 1/N$.

To estimate N, we require that $\epsilon \gg 4^{-L}$, so $N \ll 4^L$. An orbit in the symbolic space is constructed by a series of shift operations (from left to right) in the symbolic sequence. Given points in a box, an order-T partition is such that after T shifts in the symbolic sequence (or after T iterations of points in the symbolic space) these points spread out to the whole symbolic space. If a partition with N^2 boxes is well chosen, the correlation between the points of a box (encoding words) and the points of the boxes containing T forward iterations of these points (or words separated by a distance of T nucleotides) decays to approximately zero for a small finite 'time' T. For an order-T Markov partition, the correlation between initial points, and their T forward iterations decays to zero for the finite time T. Our goal is to construct an approximation of a Markov partition model to the DNA [28,29]. Because the dynamics generating the symbols are not known, our model can also be considered as a Hidden Markov model [30,31]. This model allow us to predict how groups of words are mapped to other groups of words.

In the symbolic sequences of our DNA model there are forbidden words, words that were considered in [18] to characterize the DNA, but in our symbolic space there are no forbidden groups of words. Any two groups of words is separated by the same small number $2L_n$ of nucleotides. It is the likelihood of appearance of the two separated groups of words that defines the grammatical rules of our Markovian language model.

In order to create a partition that approximately satisfies properties of a Markov system, we define the correlation (C) and the MIR.

Correlation is a powerful measure to establish relationship between two variables. We measure correlation in our symbolic space by

$$C(N, \tau) = \sum_{ij} (p_N(i)p_N(i|j)_\tau - p_N(i)p_N(j)), \tag{3.1}$$

where N is the number of rows or columns of the partition, $p_N(i)$ is probability of points being in box i and $p_N(i|j)_\tau$ is the transition probability of points going from box i to box j after τ iterations, $p_N(j)$ is the probability of being in box j. In the usual notation for the conditional probabilities, $p(i|j)_\tau$, notation adopted in this work, represents $p(j|i)_\tau$, the conditional probability of j given i. The correlation between points and their iterations measures the degree of dependence between them. Our main hypothesis to

model the DNA is to assume that there is a finite N and a finite τ for which $C(N, \tau = T) \cong 0$, i.e. the DNA has mixing properties and can be described by an approximate Markov system. Notice that if $p_N(i)p_N(i|j)_\tau - p_N(i)p_N(j) = 0$ for all i and j (strong mixing), then the partition is Markov for $\tau = T$ and generates a memoryless process, i.e. the points (δ_j, γ_j) can be considered as random uncorrelated variables if sampled every τ iterations, i.e. (δ_j, γ_j), $(\delta_{j+\tau}, \gamma_{j+\tau})$, $(\delta_{j+2\tau}, \gamma_{j+2\tau})$, ... Another way to understand the memoryless property is by noticing that if $p_N(j) = p_N(i|j)_\tau$, then the probability of being in box j does not depend on which box i the point was. For a mathematical demonstration and for a detailed discussion about the relationship among correlation, mixing, and Markov chains and partitions, see the electronic supplementary material.

The MIR [32] is given by

$$\text{MIR}(\delta, \gamma) = \frac{I_s(\delta, \gamma, N)}{T(N)}, \quad (3.2)$$

where $T(N)$ is the value such that $C(N, \tau = T) \to 0$. In practice $T(N)$ is the smallest value for which $\mathbf{A}^T$ ($\mathbf{A}$ to the power of T) has no zero elements, a necessary condition for $C(N, \tau = T) = 0$, where $\mathbf{A}$ is the transitional probability matrix with elements $p_N(i|j)_1$. I_s is the mutual information defined as

$$I_s(\delta, \gamma) = H_\delta + H_\gamma - H_{\delta;\gamma}, \quad (3.3)$$

where, $H_\delta = -\sum_i P_\delta(i) \log(P_\delta(i))$, $H_\gamma = -\sum_j P_\gamma(j) \log(P_\gamma(j))$ and $H_{\delta;\gamma} = -\sum_{i,j} P_{\delta\gamma}(i,j) \log(P_{\delta\gamma}(i,j))$, with $P_\delta(i)$ representing the probability of points in column i of the coordinate where δ is being plotted, $P_\gamma(j)$ is the probability of points in row j of the coordinate where γ is being plotted, and $P_{\delta;\gamma}(i,j)$ is the joint probability of finding points in the box (i, j) formed by the overlaps of column i with row j. Mutual information between δ and γ measures how much δ is dependent on γ and also the amount of information they share. As defined by equation (3.2), the MIR measures in average how much information per unit of 'time' (or per symbol) the past length-L symbolic sequences exchange with the future length-L symbolic sequences. Then, we check for two criteria to search for our approximate Markov partitions.

(i) *The mixing property*: $N = N^*$ and $\tau = T^*$ of the approximate Markov partition is found by minimizing the correlation $C(N, \tau)$. This criterion is needed in order for the partition to behave as an approximate order-τ Markov partition: the correlation between words separated by T^* nucleotides is close to zero. It also allows that the MIR can be calculated by equation (3.2).

(ii) *The stationary property*: $\text{MIR}(\delta, \gamma)$ obtained for $N = N^*$ is maximal and it remains invariant for any optimal value N^*. Invariance of MIR for different N reflects the fact that our Markov model is stationary with respect to the optimal values of N and that the partition is generating, in the sense that information is preserved for partitions with different order and in addition a union of 4^2 boxes of an order-T_2 partition belong to one box of an order-T_1 partition, where $T_1 < T_2$. For example, the boxes whose name are 'X0.0Y' (order-4 partition) belong to the box '0.0' (order-2 partition).

Concerning criterion 1, figure 2a shows the correlation colour coded for different grid sizes N and different τ. We observe that with the increase in τ the correlation decays very rapidly. All the partitions have τ for which $C(N, \tau) \cong 0$. In other words, we can create models of the DNA considering different word lengths and different groups. From now on, however, we focus our attention into a Markovian partition that is also stationary. To that goal, we consider criterion 2.

Concerning criterion 2, as shown in figure 2b, two peaks are observed for the MIR one at $N = 4$ and another at $N = 16$. It naturally comes from the figure that the optimal N is power of 4 (excluding the peak at 14). This is not surprising since 4^{2L_n} is the number of different words of length $2L_n$, i.e. the number of different groups of words of length $2L$. A Markov partition must have a box for each group of words. So, $N^2 = 4^{2L_n}$. From this, we conclude that $L_n = \frac{1}{2} \log_2(N)$, so boxes have symbolic names of length $2L_n$. In these cases, note that $T = 2L_n$, the time for the correlation to decay approximately to zero. Therefore, the order ϕ of a partition is defined as

$$\phi = 2L_n = \log_2(N) = T. \quad (3.4)$$

From this point on, we choose $N^* = 16$, as the resolution of our optimal partition. The selected N^* should also have the largest correlation for $\tau = 1$. In here, we want to have a Markov system for $\tau > 1$, but a strongly correlated system for $\tau = 1$, such that we can create network representations of the DNA where the edges represent the probability measure of points going from box i to box j group of words that
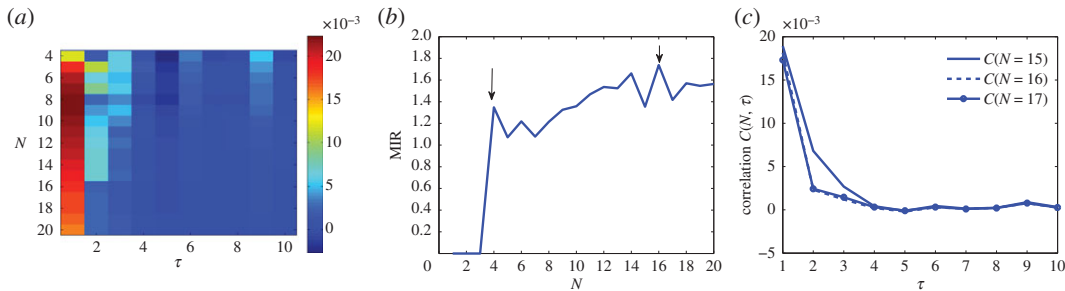
Figure 2. Correlation and MIR for determining the resolution of the partition, which in turn provides the grouping of words. (*a*) Correlation for different partitions, as a function of N and τ. (*b*) MIR as a function of N for $L = 4$. Maximization of MIR is attained for $N = 16$. The arrow shows that maximization is possible for specific N. The N should be a multiple of 4^n. (*c*) Correlation values for $N = 15$, $N = 16$ and $N = 17$ with respect to τ. $C(N, \tau) \cong 0$ at around $\tau \geq 4$.

are one nucleotide apart. In figure 2c, we show the correlation decay for the partitions with $N = 15, 16$ and 17. In all cases, the correlation is large for $\tau = 1$ and it decays fast.

It is known that informational quantities calculated in spaces with finite resolution and with a finite number of points lead to overestimation of these quantities [33,34]. The reason being that the probabilities calculated in a non-Markov partition would carry spurious correlations. Since our probabilistic quantities are being calculated over partitions that are approximately Markov, informational quantities obtained from them should be minimizing the appearance of overestimations.

3.3. The Markovian model of the DNA

Each box has a name given by a symbolic sequence of length $2L_n$. Every point belonging to that box will have a symbolic sequence that contains the name of the box and therefore, up to T iterations, we are able to have partial information about the evolution of the trajectory in the symbolic space. Then, for $\tau = T = \phi$ iterations, the systems becomes memoryless and the transition matrix $\mathbf{P}_\tau$ represents our Markov model of the DNA. Its main properties are as follows:

  (i) $C(N, \tau = \phi) \cong 0$,
 (ii) $\mathrm{MIR}(\delta, \gamma)$ is maximal and invariant over different ϕ orders, and
 (iii) $p(i|j)_\phi \cong p(j)$.

Condition (i) implies that the order—ϕ partition provides a model that has weak mixing properties. Condition (ii) implies that we search for partitions of different orders whose content of information is invariant. Condition (iii) implies that the model also behaves as a Markov system and in addition the measure provided by the order-T partition is invariant.

In the symbolic sequences of the DNA model there are forbidden words, but there are no forbidden group words. Any two groups of words is separated by the same small number $2L_n$ of nucleotides. It is the likelihood of appearance of the two separated groups of words that defines the grammatical rules of the language model.

Because the quantities considered to model the DNA are all small and finite, computational cost can be reduced. If the Markovian property is fully verified in the DNA, or at least in a piece of it, the whole information of it can be stored in the matrix $p(i|j)_\phi$, which has dimensionality $4^{2\phi}$.

4. Results and discussion

4.1. Functional network of the genome

Let us define that $p(i)$ represents the probability of being in box i in a partition of order $\phi = 4$ (the sub index N in the notation for probabilities is dropped since we set $N = 16$) and $p(i|j)_1 = n_t(i \to j)/n_t(i)$ (where $n_t(i \to j)$ represents the number of points in box i that goes to box j, and $n_t(i)$ represents the number of points in box i) is a term of the transition probability matrix, representing the transition probability of going from box i to box j, after one time iteration. A transition matrix represents a square matrix with

positive entries such that $\sum_j p(i|j)_1 = 1$ for all i. The amount of probability measure that goes from box i to box j after one time iteration is given by $p(i)p(i|j)_1$. To create a graph representing how points in the two boxes in the partition are correlated, we consider the quantity $p(i)p(i|j)_1$. An edge connecting the node i to the node j is considered to exist if $p(i)p(i|j)_1 \geq t^*$. There will be $N_E = 4.N^2$ edges, since every node (in a partition with N^2 boxes) represents groups of words that all contain the same sequence of length $2L_n = \log_2(N)$ and shifting words in box i one nucleotide (or after one iteration), results in another group of words that are encoded by symbols whose encoding occupy certainly four other boxes (see illustrations in figure 1c). There will be $N_V = N^2$ nodes in the network, and therefore the number of edges are given by $N_E = 4.N^2$.

4.2. Threshold network and associated measures

Along the lines of the work of [22], we apply our model to create a network representation of the DNA. In contrast to that work, our words do not necessarily have a biological meaning as the motifs considered in that work. The advantage of our approach is that being a Markov model, we can in principle predict very long sequence of symbols. Preserving the connections responsible for most of the information of the DNA, we create a reduced network model of the DNA, which capture the most relevant features of it, and is able to identify hub groups of words, which have a similar function to the core words identified by Gerlach & Altmann [23] in natural languages.

The informational hubs of the DNA symbolic network are found by removing nodes and their edges that transfer little measure. Ignoring the smaller $p(i)p(i|j)_1$ values by thresholding create network representations of the DNA that contains less information about it. But selecting boxes where $p(i)p(i|j)_1$ are large have the effect of revealing a network of very likely transitions between words, the informational hub networks.

Removing the values from the transition matrix can change its properties entirely and it would no longer be a transition matrix which means that quantities based on probabilities such as I_s, MIR and $C(N = 16, \tau = 1)$ cannot be calculated. The model would lose its Markov properties. So, to restore the properties of the transition matrix and to maintain the Markov properties of the model for a reduced network of groups of words, we have to rescale it to $\tilde{p}(i|j)_1$ so that the properties of a transition matrix are maintained, i.e. $\sum_j \tilde{p}(i|j)_1 = 1$.

Let the original transition matrix for $\tau = 1$ be $\mathbf{A}$ with elements $\mathbf{A}_{ij} = p(i|j)_1$ and the new thresholded matrix be $\tilde{\mathbf{A}}$. Then we obtain $\tilde{\mathbf{A}}$ by

$$\tilde{\mathbf{A}}_{ij} = \left(\frac{\mathbf{A}'_{ij}}{\sum_j \mathbf{A}'_{ij}} \right), \tag{4.1}$$

where, $\mathbf{A}'_{ij} = \mathbf{A}_{ij}$ if $\mathbf{A}_{ij} > t^*$ and $\mathbf{A}'_{ij} = 0$, otherwise. The rescaled network will therefore have $\tilde{N}_V$ nodes. The density E_d of the edges connecting two nodes representing a symbolic transition between two group of words in the rescaled network is given by

$$E_d = \frac{\tilde{N}_E}{\tilde{N}_V}, \tag{4.2}$$

where $\tilde{N}_E$ is the number of edges of the rescaled network, i.e. the number of times that $p(i)p(i|j) \geq t^*$.

Total outgoing measure of the non-thresholded network can be calculated by

$$M = \sum_{ij} p(i)p(i|j)_1. \tag{4.3}$$

The total measure is calculated by equation (4.3), but neglecting in the summation i and j for which $p(i)p(i|j)_1 \leq t^*$, we denoted by $\tilde{M}$. It lies between $[0, 1]$, where 0 means that network represents 0% of the transitions and the measure associated with the symbolic space; 1 means that the network is based on all the observed transitions.

We must also reconstruct the invariant measure based on $\tilde{\mathbf{A}}$. Given $\tilde{\mathbf{A}}$ and assuming it to be a regular matrix we want to calculate the invariant measure

$$\tilde{\mathbf{P}}(\tilde{N}_V) = \tilde{p}(i), \quad i = 1, \dots, \tilde{N}_V, \tag{4.4}$$

such that $\sum_{i=1}^{\tilde{N}_V} \tilde{p}(i) = 1$. Here we assume that the nodes of the thresholded network represent groups of words where probability of appearance is provided by a Markov memoryless process. This implies that

if $\tilde{\mathbf{P}}_0$ represents a vector of initial random values with $\tilde{N}_V$ entries forming a vector, then $\tilde{\mathbf{P}}$ with elements $\tilde{p}(i)$ is calculated by

$$\lim_{n\to\infty} \tilde{\mathbf{P}}^n = \tilde{\mathbf{P}}_0 \tilde{\mathbf{A}}^n, \tag{4.5}$$

In order to characterize a system whose probabilities and transition probabilities are given by $\tilde{p}(i)$ and $\tilde{p}(i|j)_1$, respectively, we calculate the MIR, which measures exchange of information per symbol between two groups of words of length $2L$ shifted one nucleotide apart defined as

$$\widetilde{\mathrm{MIR}}(\tilde{N}_V, t^*) = \frac{I_s(\tilde{N}_V, t^*)}{\tilde{T}}, \tag{4.6}$$

where I_s is defined as

$$I_s(\tilde{N}_V, t^*) = \sum_{i,j} \tilde{p}(i)\tilde{\mathbf{A}}_{ij} \log\left(\frac{\tilde{p}(i)\tilde{\mathbf{A}}_{ij}}{\tilde{p}(i)\tilde{p}(j)}\right). \tag{4.7}$$

Notice that $\tilde{p}(i)\tilde{A}_{ij}$ is a joint entropy. The average rate of information contained in all groups of words is calculated by Shannon's entropy rate:

$$S(\tilde{N}_V, t^*) = \frac{S_n(\tilde{N}_V, t^*)}{\tilde{T}}, \tag{4.8}$$

where S_n is defined as

$$S_n(\tilde{N}_V, t^*) = \sum_i \left(\tilde{p}(i) \log \frac{1}{\tilde{p}(i)}\right). \tag{4.9}$$

$\tilde{T}$ is the time for which $\tilde{\mathbf{A}}^{\tilde{T}}$ has only non-null elements. If the network is not connected (it can be decomposed in two or more sub-networks), then $\tilde{T}$ is the average of all $\tilde{T}$ for each sub-network. To determine how the mutual information changes when we remove the nodes, we calculate the mutual information I_s between two groups of words per node $\tilde{N}_V$ for each threshold t^*:

$$\sigma_{\mathrm{I}} = \frac{I_s(\tilde{N}_V, t^*)}{\tilde{N}_V}. \tag{4.10}$$

Similarly, we calculate the Shannon entropy per node also, defined by

$$\sigma_{\mathrm{S}} = \frac{S_n(\tilde{N}_V, t^*)}{\tilde{N}_V}. \tag{4.11}$$

Figure 3*a* shows different informational quantities and their behaviour as we change the threshold t^*. The knock-out of only one group of words for $t^* = 0.03$ leads to an approximate 20% of reduction in the information produced (S) and exchanged by the network. When compared with the decrease in the number of nodes at this t^*, this number changes only by 1 but the change in $\widetilde{\mathrm{MIR}}$ and S shows evident loss of information. For $t^* \in [0.1, 0.13]$, $\widetilde{\mathrm{MIR}}$ and S remain constant and decreases for $t^* \cong 0.14$. This pattern of $\widetilde{\mathrm{MIR}}$ and S of being constant and then decreasing goes until $t^* = 0.24$. The picture shows three changing steps for $\widetilde{\mathrm{MIR}}$ and S for $t^* \leq 0.18$. Each step represents a single node in the thresholded network being eliminated. The nodes lost, respectively, for these three steps represent the 'CTAG', 'CCTG' and 'TTAG' group of words. These words belong to the well-defined group-I and -II tetramer classes of nucleotides [35]. Studies have shown that these groups of tetramers are actively involved in base mismatch repair in *E. coli* and are known as very short patch [36]. The word 'CTAG' is a well-known palindromic sequence which is rarely present in protein-coding regions but is abundantly present in genes coding for structural RNAs [37]. Almost $\frac{2}{5}$ of the information of the DNA is contained in symbolic sequences formed by these three length-4 words. A smooth decay in $\tilde{M}$ and $\tilde{N}_V/N^2$ is observed for $t^* \in [0.12, 0.20]$ but suddenly an abrupt change is noticed in $\tilde{M}$ and $\tilde{N}_V/N^2$ for the interval $t^* \in [0.2, 0.25]$. The number of nodes of the thresholded network change from 256 at $t^* = 0$ to 85 at $t^* = 0.25$ after this abrupt decay. Although the percentage of nodes remaining is just around 33% of all possible nodes, the thresholded network's nodes contain words which appear in most of the genes and $\tilde{M}$, $\widetilde{\mathrm{MIR}}$ and S are moderately high. Moving further to $t^* \in [0.25, 0.3]$, $\widetilde{\mathrm{MIR}}$ and S increases for $t^* \in [0.25, 0.27]$ and then decays abruptly for $t^* \in [0.27, 0.3]$. The number of nodes of the rescaled network for $t^* \in [0.25, 0.3]$ change from 85 at $t^* = 0.25$ to 5 at $t^* = 0.30$. For $t^* \geq 0.3$, the number of nodes remains constant, and not only $S = \widetilde{\mathrm{MIR}}$ but both quantities are very high. This shows that the minimum number of nodes that remain even after thresholding the data is 5.

The fact that $S = \widetilde{\mathrm{MIR}}$ means that all the measurements of a box is mapped to another unique box, and consequently in equation (4.7), $\tilde{p}(i)\tilde{\mathbf{A}}_{ij} = \tilde{p}(j)$. Groups of words map uniquely to another group and every
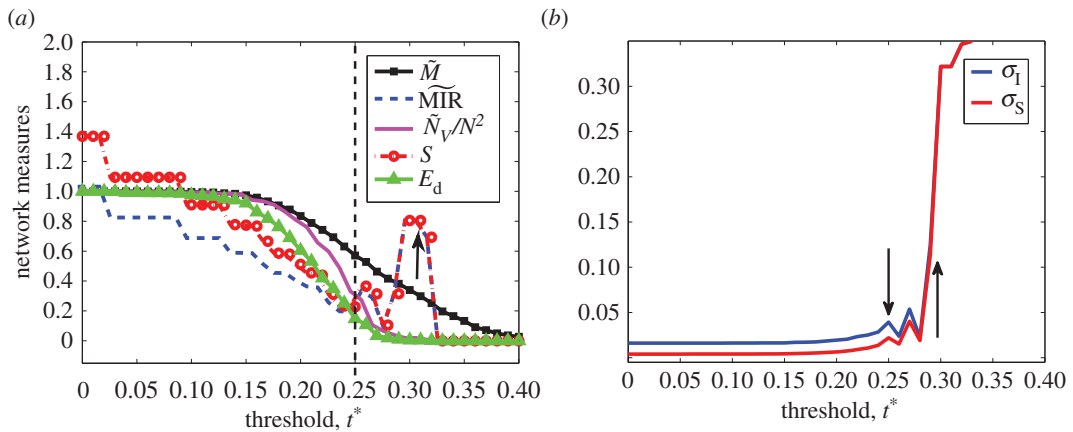
Figure 3. Characterization of a network representation of DNA based on information theoretical quantities as a function of the threshold t^*. (*a*) $\tilde{M}$ is the total probabilistic measure of the thresholded network (black line and square), MIR represents blue dashed line, the pink line represents the relative number of nodes $\tilde{N}_V/N^2$, S is the Shannon entropy rate (red line and circles) and E_d is the density of edges (green line and triangles). The dashed line represents the $t^* = 0.25$. (*b*) The curves show how information grows with the decrease in the number of nodes but retaining only nodes with high information content. σ_I is mutual information per node and σ_S is Shannon entropy per node. The arrows represent the unique threshold points where the information changes.

node of the network has a degree 1. To make this argument rigorous, if $\tilde{p}(i)\tilde{\mathbf{A}}_{ij} = \tilde{p}(i)$, it means that $\tilde{\mathbf{A}}_{ij} = 1$ for that i. But $\sum_j(\tilde{\mathbf{A}}_{ij}) = 1$ which means that $\tilde{\mathbf{A}}_{ij} = 0$ for $i \neq j$. So, we can write equation (4.7) as

$$\tilde{I}_s(\tilde{N}_V, t^*) = \sum_i \sum_{j=i} \tilde{p}(i) \log\left(\frac{\tilde{p}(i)}{\tilde{p}(i)\tilde{p}(j)}\right). \tag{4.12}$$

and since $j = i$ in the summation of equations (4.12) and (4.7) can be written as

$$\tilde{I}_s(\tilde{N}_V, t^*) = \sum_i \tilde{p}(i) \log\left(\frac{1}{\tilde{p}(i)}\right). \tag{4.13}$$

Therefore,

$$\tilde{I}_s(\tilde{N}_V, t^*) = S_H(\tilde{N}_V, t^*). \tag{4.14}$$

In figure 3*b*, we show how $\widetilde{\text{MIR}}$ and S increase per $\tilde{N}_V$. Both curves are roughly constant for $t^* \in [0, 0.18]$. For larger t^* values, these quantities start to fluctuate. A small peak is observed at $t^* = 0.25$ for both quantities while a few fluctuations happen for the range of $t^* \in [0.25, 0.29]$ but then these quantities abruptly increase at $t^* = 0.30$. This is not surprising, since $\tilde{I}_s$ maximizes when $\tilde{p}(i)\tilde{\mathbf{A}}_{ij} = \tilde{p}(j)$. These results suggest that the networks obtained for the interval $t^* \in [0.25, 0.30]$ represent groups of words that exchange high amounts of information (relative per remaining nodes). Therefore, if a word is found in the DNA that belongs to one of the groups of these rescaled networks the predictability of the possible iterated word is very high.

In figure 4, we show two examples of the rescaled networks obtained. The smallest network obtained for $t^* = 0.30$ with five nodes is shown in figure 4*a*. The sequence generated from the cycle of length three connecting three nodes is 'GCAGCA'. This sequence is formed by a likely transition of words of four symbols that contain high amounts of information. Therefore, one should expect the appearance of this sequence in the DNA of *E. coli*. One should also expect these transitions between the length-4 words. From figure 3*a*, one sees that this network at $t^* = 0.3$ has values of MIR and S comparable to the larger network at 85 nodes (at $t^* = 0.25$). From figure 3*b*, it is clear that the information content per node of this network is one of the highest.

In figure 4*b*, we show a network of 85 nodes. This network contains only around 33% of all possible nodes, but surprisingly at $t^* = 0.25$, $\tilde{M}$ is about half the total measure of the full network, which means that these words are very likely in the genome. Remarkably, this network is the first smallest network that can generate a large sequence of words with any periodicity, since it has cycles that are connected and that forms words which appear with periodicity of 1, 2 and 3. This network reminds us of the Sharkovskii's theorem which states that if a dynamical system with some properties has an equilibrium point, a period 2 and a period 3 orbit then it must have periodic points for every other period. This network is very special. From figure 3*a*, we conclude that adding just few more edges in this network (decreasing from t^*
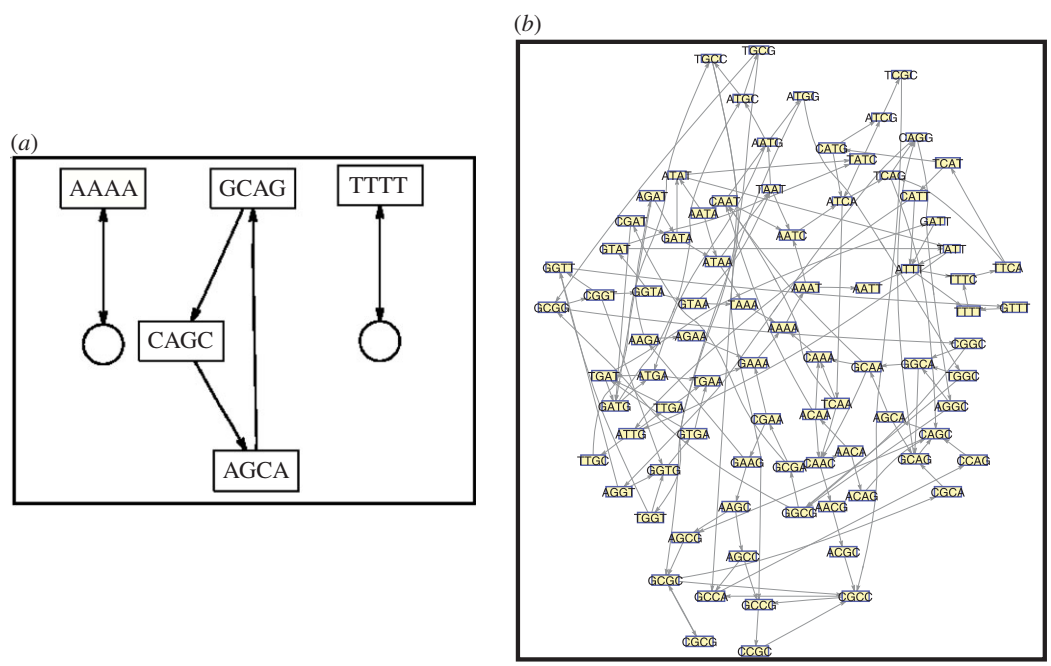
Figure 4. (*a*) Network with five nodes at $t^* = 0.3$. In the centre, the words GCAC, CAGC, AGCA are connected in a cycle of length 3. The rectangular boxes represent the nodes and the circles are used to represent a self-loop. (*b*) Network with 85 nodes at $t^* = 0.25$.

from 0.25), we would restore almost all the information content of *E. coli* DNA. If only a few more nodes are removed (increasing t^* from 0.25) we see that information (I_s and S) abruptly decays. This network therefore should represent the structural 'skeleton' of *E. coli*, i.e. the sequences generated by it should be fundamental to *E. coli*. The analogy of the network structure possessing a skeleton property reflects the idea that a skeleton provides the structural stability of a configuration. Removing a piece of it leads to an unstable configuration. Introducing more structure, the configuration becomes stronger and more stable. To understand the importance of networks shown in figure 4, we created a genomic sequence from the transitions in the group of words depicted by the networks and matched them with the known genes of *E. coli*. Some of the genes which contained sequence from both the networks shown in figure 4 are *nrfA*, *flgF*, *mnmC*, *cysM*, *xseA*, and *hscB*. These genes are known to be involved in the structural foundation of the organisms by either being a structural gene or producing proteins for structural stability. They are also involved in DNA-binding and promoting stress-induced mutagenesis. [38–40]. We also found *iscS*, *iscU* and *iscR* which help in DNA-binding and scaffold protein for iron–sulfur cluster assembly. These three genes work along with *hscB* and are involved in some pathways like alanine biosynthesis III, molybdenum cofactor biosynthesis and thiazole biosynthesis I [41].

5. Validation of the model and predictability of group of words for genes

To validate the model and study its predictability, we transform $\tilde{\mathbf{A}}(t^*)$ from the whole genome to an adjacency matrix $\mathbf{G}^{E.\,coli}(t^*)$, with elements $G_{ij}^{E.\,coli}(t^*) = 1$ if the group of words in box i iterate to box j or $G_{ij}^{E.\,coli}(t^*) = 0$ if there are no transitions from box i to box j. Then, we evaluate the efficiency of this model at different levels of t^*, by studying its ability to predict genes g_i at t^* levels. We create a symbolic sequence for each g_i, similar to the method we followed for genomic sequence ($N = 16, L = 4$), and obtain an adjacency matrix $\mathbf{G}(g_i, t^*)$ from the transition matrix $\mathbf{B}(g_i, t^*)$ for gene i. $G_{ij}(g_i, t^*) = 1$ if the group of words where points are in a box i are iterated to box j (so $B_{ij}(g_i, t^*) > 0$, and $G_{ij}(g_i, t^*) = 0$, if $i = j$ or $B(g_i, t^*) = 0$, otherwise). We then compare $\mathbf{G}(g_i, t^*)$ with $\mathbf{G}^{E.\,coli}(t^*)$, the adjacency matrix for the genome of *E. coli* defined at different threshold t^* and see how efficiently we can predict the existence of groups of words and their transitions in each gene. For this purpose, we calculate parameters for the True Positive Rate (TPR) or sensitivity S_n, and False Positive Rate (FPR) or specificity S_p of the model. These parameters are defined as

$$S_n = \frac{T_P}{T_P + F_N} \tag{5.1}$$
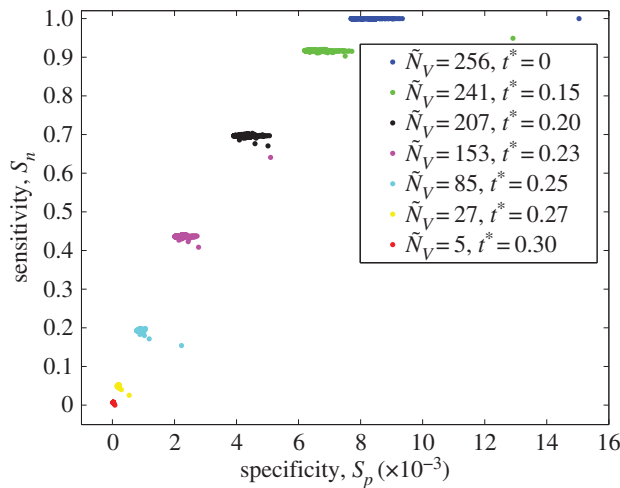
Figure 5. Specificity S_p shown on the horizontal axis and sensitivity S_n shown on the vertical axis plot for *E. coli* with different levels of threshold t^*.

and

$$S_p = \frac{T_N}{T_N + F_P},$$ (5.2)

where T_P is the number of transitions between two groups of words correctly predicted, therefore, T_P for gene g_i is defined as $T_P^{E.\,coli}(g_i, t^*) = \sum_{ij}^{1,1}(G_{ij}^{E.\,coli}(t^*) - G_{ij}(g_i, t^*))$, where we only take into consideration all the i and j values of $G_{ij}^{E.\,coli}(t^*)$ and $G_{ij}(g_i, t^*)$ which are equal to 1. The symbol $\sum_{ij}^{1,1}$ represents a summation that is only carried out when the variables inside the argument are equal to the super index. F_N is the number of words that were wrongly predicted, $F_N^{E.\,coli}(g_i, t^*) = \sum_{ij}^{1,0}(G_{ij}^{E.\,coli}(t^*) - G_{ij}(g_i, t^*))$, this can happen only when $G_{ij}^{E.\,coli}(t^*) = 1$ and $G_{ij}(g_i, t^*) = 0$, meaning that a transition from the group of words in box i are mapped to box j are not present but have been wrongly predicted by the model. $T_N^{E.\,coli}(g_i, t^*) = \sum_{ij}^{0,0}(G_{ij}^{E.\,coli}(t^*) - G_{ij}(g_i, t^*))$, in this case we consider all the values of $G_{ij}^{E.\,coli}(t^*)$ and $G_{ij}(g_i, t^*)$ that are equal to zero, meaning that the a transition from the group of words in box i are mapped to box j do not exist and the model also does not predicts them. $F_P^{E.\,coli}(g_i, t^*) = \sum_{ij}^{0,1}(G_{ij}^{E.\,coli}(t^*) - G_{ij}(g_i, t^*))$, happens when $G_{ij}^{E.\,coli}(t^*) = 0$ but $G_{ij}(g_i, t^*) = 1$.

Figure 5 shows the $S_n^{E.\,coli}(t^*)$ versus $S_p^{E.\,coli}(t^*)$ plot for the different thresholds t^* considered to construct $\mathbf{G}^{E.\,coli}(t^*)$ and $\mathbf{G}(g_i, t^*)$. The results obtained for each t^* network has been represented with different colours and symbols. The results obtained for the network composed of 256 nodes ($t^* = 0$) is shown with blue points. The $S_n^{E.\,coli}(t^*)$ value for this network is $\simeq 1$ and $S_p^{E.\,coli}(t^*) \cong 0$, meaning that this network correctly predicts the transition of groups of words for genes. Having the TPR or S_n constant for all genes at a given threshold shows that our model can predict every gene with similar accuracy. Regardless of the fact that the model was constructed from the whole-genome data, we could still predict transition of groups of words in each gene. Notice that how words are mapped to words is not relevant in our model, but how groups of words are mapped to other groups of words. This is a consequence of the Markovian characteristics of the model that the property of the 'whole' reflects also the property of the 'parts'.

In the electronic supplementary material, we show how our model can be used to detect similarity between genes of different organisms and shows that our model predicts better the appearance of groups of words then standard probabilistic methods that predicts the appearance of particular words.

6. Data collection

All the genomic sequences for organisms were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/). The NCBI Reference Sequence ID for each organism are as follows: (*E. coli*: GenBank:NC_000913.3, *Sh. dysenteriae*: GenBank:NC_007606, *R. fascians*: GenBank:NC_021080.1 and *Sa. cerevisiae*: GenBank:NC_001133.9).

7. Conclusion

This work proposes a Markovian language model for the DNA. We encode nucleotide sequences into symbolic sequences of finite length, regarded as words and then encode these words into points belonging to a two-dimensional symbolic space of the DNA; from which we establish the functional connectivity between any two regions in this symbolic space, representing two groups of similar words, i.e. the likelihood of having a word belonging to a group being followed after some nucleotides into another word that belongs to another group of words.

We construct a Markov network representation of the DNA, the nodes representing groups of words and their probabilities and the edges their transition probabilities, such that the statistics of the transition between groups of words is approximately memoryless. Our model allows for a reduction in the complexity of the DNA. For the *E. coli*, we showed that a network of only 85 nodes (representing 85 groups of words) contains most of the information of its DNA, composed of more than 1 billion nucleotides. On the other hand, a network with all but three words looses almost $\frac{2}{5}$ of its information content. We have also shown that our model can be used as a similarity measure to detect similar symbolic genes in different organisms (see the electronic supplementary material). The results demonstrated that genes with not entirely known function were similar to genes with well-established and known function as in case of *E. coli* and *Sa. cerevisiae*.

We have shown that for the DNA, an approximate Markov partition can be constructed assuming equal-sized cells. Among all possible sizes, we have shown that our approximate Markov partition exists when there are 4^L cells. Our approach can be extended to other systems as long as a Markov partition is obtained. To construct Markov partitions for systems presenting time-varying delays or partially accessible information, one could consider the works in [42,43], or [29] for stochastic systems, or the work in [44] for low-dimensional dynamical systems.

This model can also be used to calculate the statistics of recurrence of groups of words in DNA (see the electronic supplementary material). Studying the recurrence of these groups of words can be used in determining the stochastic and the deterministic nature of the DNA which contributes to the genes, coding and non-coding regions of the DNA, and provide an explanation for the evolution of the DNA from simple organisms to the Human genome in terms of how the recurrence of the DNA has become more or less memoryless. Our model allows us to deduce analytical expressions for the probability density of returns of groups of words. The more Markovian a piece of the DNA is, the more accurate are our analytically obtained density. In a publication to be submitted elsewhere, we will demonstrate that the coding part of the DNA is more Markovian (random) than the non-coding, and the sequences appearing on it can have their short- and long-term behaviour well predicted by our model.

References

1. Nelson DL, Cox MM. 2010 *Lehninger principles of biochemistry*, 5th edn. New York, NY: W. H. Freeman.

2. Kim S, Tang H, Mardis ER. 2007 *Genome sequencing technology and algorithms*, 1st edn. Norwood, MA: Artech House Inc.

3. Berger AL, Della Pietra VJ, Della Pietra SA. 1996 A maximum entropy approach to natural language processing. *Comput. Linguist.* **22**, 39–71.

4. Lewis D, Jones KS. 1996 Natural language processing for information retrieval. *Commun. ACM* **39**, 92–101. (doi:10.1145/234173.234210)

5. Harispe S, Ranwez S, Janaqi S, Montmain J. 2015 *Semantic similarity from natural language and ontology analysis* (*synthesis Lectures on Human Language Technologies.*) San Rafael, CA: Morgan & Claypool Publishers.

6. Richards RJ. 1989 Darwin and the emergence of evolutionary theories of mind and behaviour. *J. Hist.* *Behav. Sci.* **25**, 302–302. (doi:10.1002/1520-6696(19 8907)25:3<302::AID-JHBS2300250346>3.0.CO;2-D)

7. Darwin C. 1916 *The expression of the emotions in man and animals*. New York, NY: Appleton and Co.

8. Botstein D, Cherry MJ. 1997 Molecular linguistics: extracting information from gene and protein sequences. *Proc. Natl Acad. Sci. USA* **94**, 5506–5507. (doi:10.1073/pnas.94.11.5506)

9. Brendel V, Beckmann JS, Trifonov EN. 1986 Linguistics of nucleotide sequences: morphology

and comparison of vocabularies. *J. Biomol. Struct. Dyn.* **4**, 11–21. (doi:10.1080/07391102.1986.10507643)

10. Gramatikoff K, Boyd SE, Smith JW, Keith JM. 2014 Genomic taxonomy boost by lexical clustering. *J. Invest. Genomics* **1**, 1–15.

11. Afreixo V, Bastos CA, Pinho AJ, Garcia SP, Ferreira PJ. 2009 Genome analysis with inter-nucleotide distances. *Bioinformatics* (*Oxford, England*) **25**, 3064–3070. (doi:10.1093/bioinformatics/btp546)

12. Anastassiou D. 2001 Genomic signal processing. *Signal Process. Mag. IEEE* **18**, 8–20. (doi:10.1109/79.939833)

13. Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Matsa ME, Peng CK, Simons M, Stanley HE. 1995 Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Phys. Rev. E* **51**, 5084–5091. (doi:10.1103/PhysRevE.51.5084)

14. Cristea PD. 2003 Large scale features in DNA genomic signals. *Signal Process.* **83**, 871–888. (doi:10.1016/S0165-1684(02)00477-2)

15. Voss RF. 1992 Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.* **68**, 3805–3808. (doi:10.1103/PhysRevLett.68.3805)

16. Nair ASS, Mahalakshmi T. 2005 Visualization of genomic data using inter-nucleotide distance signals. *Proc. IEEE Genomic Signal Process*.

17. Baptista MS, Caldas IL, Baptista MS, Baptista CS, Ferreira AA, Heller MV. 2000 Low-dimensional dynamics in observables from complex and higher-dimensional systems. *Physica A* **287**, 91–99. (doi:10.1016/S0378-4371(00)00448-9)

18. Hao B, Xie H. 2009 Factorizable language: from dynamics to biology. In *Reviews of nonlinear dynamics and complexity*, pp. 147–186. University of Kiel, Germany: Wiley-VCH Verlag GmbH & Co. KGaA.

19. Arneodo A, Vaillant C, Audit B, Argoul F, d'Aubenton Carafa Y, Thermes C. 2011 Multi-scale coding of genomic information: from DNA sequence to genome structure and function. *Phys. Rep.* **498**, 45–188. (doi:10.1016/j.physrep.2010.10.001)

20. Anishchenko VS, Vadivasova TE, Okrokvertskhov GA, Strelkova GI. 2003 Correlation analysis of dynamical chaos. *Physica A* **325**, 199–212. (doi:10.1016/S0378-4371(03)00199-7)

21. Feldman DP. 2012 *Chaos and fractals: an elementary introduction*. New York, NY: Oxford University Press.

22. Sinatra R, Condorelli D, Latora V. 2010 Networks of motifs from sequences of symbols. *Phys. Rev. Lett.* **105**, 178702. (doi:10.1103/PhysRevLett.105.178702)

23. Gerlach M, Altmann EG. 2013 Stochastic model for the vocabulary growth in natural languages. *Phys. Rev. X* **3**, 021006. (doi:10.1103/PhysRevX.3.021006)

24. Kalimeri M, Constantoudis V, Papadimitriou C, Karamanos K, Diakonos FK, Papageorgiou H. 2012 Entropy analysis of word-length series of natural language texts: effects of text language and genre. *Int. J. Bifurc. Chaos* **22**, 1250223. (doi:10.1142/S0218127412502239)

25. Douglas L, Brian M. 1995 *An introduction to symbolic dynamics and coding*. Cambridge, UK: Cambridge University Press.

26. Crutchfield JP, Packard NH. 1982 Symbolic dynamics of one-dimensional maps: entropies, finite precision, and noise. *Int. J. Theoret. Phys.* **21**, 443–466. (doi:10.1007/BF02650178)

27. Rubino G, Sericola B. 2014 *Markov chains and dependability theory*. Cambridge, UK: Cambridge University Press.

28. Bollt E, Skufca J. 2005 *Markov partitions, encyclopedia of nonlinear science*. Routledge, New York: Alwyn Scott.

29. Bollt E, Góra P, Ostruszka A, Życzkowski K. 2008 Basis markov partitions and transition matrices for stochastic systems. *SIAM J. Appl. Dyn. Syst.* **7**, 341–360. (doi:10.1137/070686111)

30. Rabiner LR. 1989 A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286. (doi:10.1109/5.18626)

31. Yano M, Kato Y. 2014 Using hidden Markov models to investigate G-quadruplex motifs in genomic sequences. *BMC Genomics* **15**, S15. (doi:10.1186/1471-2164-15-S9-S15)

32. Baptista MS, Rubinger RM, Viana ER, Sartorelli JC, Parlitz U, Grebogi C. 2012 Mutual information rate and bounds for it. *PLoS ONE* **7**, e46745. (doi:10.1371/journal.pone.0046745)

33. Steuer R, Kurths J, Daub CO, Weise J, Selbig J. 2002 The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* **18**, S231–S240. (doi:10.1093/bioinformatics/18.suppl_2.S231)

34. Herzel H, Schmitt AO, Ebeling W. 1994 Finite sample effects in sequence analysis. *Chaos, Solitons Fractals* **4**, 97–113. (doi:10.1016/0960-0779(94)90020-5)

35. Gutiérrez G, Casadesús J, Oliver J, Marine A. 1994 Compositional heterogeneity of the *Escherichia coli* genome: a role for VSP repair? *J. Mol. Evol.* **39**, 340–346. (doi:10.1007/BF00160266)

36. Bhagwat AS, McClelland M. 1992 DNA mismatch correction by Very Short Patch repair may have altered the abundance of oligonucleotides in the *E. coli* genome. *Nucleic Acids Res.* **20**, 1663–1668. (doi:10.1093/nar/20.7.1663)

37. Blattner FR *et al.* 1997 The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462. (doi:10.1126/science.277.5331.1453)

38. Eaves DJ, Grove J, Staudenmann W, James P, Poole RK, White SA, Griffiths L, Cole JA. 1998 Involvement of products of the nrfEFG genes in the covalent attachment of haem c to a novel cysteineâĂŞlysine motif in the cytochrome c552 nitrite reductase from *Escherichia coli*. *Mol. Microbiol.* **28**, 205–216. (doi:10.1046/j.1365-2958.1998.00792.x)

39. Jones CJ, Macnab RM, Okino H, Aizawa S-I. 1990 Stoichiometric analysis of the flagellar hook-(basal-body) complex of *Salmonella typhimurium*. *J. Mol. Biol.* **212**, 377–387. (doi:10.1016/0022-2836(90)90132-6)

40. Al Mamun AA *et al.* 2012 Identity and function of a large gene network underlying mutagenic repair of DNA breaks. *Science* **338**, 1344–1348. (doi:10.1126/science.1226683)

41. Keseler IM *et al.* 2013 Ecocyc: fusing model organism databases with systems biology. *Nucleic Acids Res.* **41**, D605–D612. (doi:10.1093/nar/gks1027)

42. Wei Y, Qiu J, Karimi HR, Wang M. 2015 A new design of H_∞ filtering for continuous-time Markovian jump systems with time-varying delay and partially accessible mode information. *Signal Process.* **93**, 2392–2407. (doi:10.1016/j.sigpro.2013.02.014)

43. Wei Y, Peng X, Qiu J, Jia S. 2015 H_∞ filtering for two-dimensional continuous-time Markovian jump systems with deficient transition descriptions. *Neurocomputing* **167**, 406–417. (doi:10.1016/j.neucom.2015.04.054)

44. Hirata Y, Aihara K. 2013 Estimating optimal partitions for stochastic complex systems. *Eur. Phys. J. Spec. Top.* **222**, 303–3015. (doi:10.1140/epjst/e2013-01843-x)