Symbolic complexity for nucleotide sequences: a sign of the genome structure

R Salgado-García and E Ugalde

1 Centro de Investigación en Ciencias-IICBA, Universidad Autónoma del Estado de Morelos, Avenida Universidad 1001, Cuernavaca Morelos, C.P. 62209, Mexico
2 Instituto de Física, Universidad Autónoma de San Luis Potosí, Avenida Manuel Nava 6, Zona Universitaria, 78290 San Luis Potosí, Mexico

E-mail: raulsg@uaem.mx

Received 18 January 2016, revised 28 August 2016
Accepted for publication 5 September 2016
Published 10 October 2016

Abstract

We introduce a method for estimating the complexity function (which counts the number of observable words of a given length) of a finite symbolic sequence, which we use to estimate the complexity function of coding DNA sequences for several species of the Hominidae family. In all cases, the obtained symbolic complexities show the same characteristic behavior: exponential growth for small word lengths, followed by linear growth for larger word lengths. The symbolic complexities of the species we consider exhibit a systematic trend in correspondence with the phylogenetic tree. Using our method, we estimate the complexity function of sequences obtained by some known evolution models, and in some cases we observe the characteristic exponential-linear growth of the Hominidae coding DNA complexity. Analysis of the symbolic complexity of sequences obtained from a specific evolution model points to the following conclusion: linear growth arises from the random duplication of large segments during the evolution of the genome, while the decrease in the overall complexity from one species to another is due to a difference in the speed of accumulation of point mutations.

Keywords: symbolic complexity, symbolic dynamics, genome structure

(Some figures may appear in colour only in the online journal)

1. Introduction

Nucleotide sequences carry information about an individual’s life as well as their evolutionary history. A lasting problem in modern genomics is how this information is organized in the genome or, in other words, what the structure of the genome is from an information theory
point of view. In the past decades, several pieces of work have addressed the answer to this question. Nowadays, some universal characteristics have been identified, such as the 3-base and 10.5-base periodicities \[1\], the long-range correlations \[2–4\], and the fractal structure and the scaling properties of the DNA \[5–7\], among others. Understanding the underlying processes leading to the current structure of the genome is still an open problem. This has resulted in the introduction of several models of genome evolution in an attempt to reproduce the above-mentioned characteristics \[8–13\]. Knowing how this structure emerges could result in a deeper understanding of the physics of DNA (such as protein diffusion along the DNA \[14–17\], its translocation properties through a nanopore \[18\] or the condensation-decondensation process of the nucleosome \[19\]—or even in applications regarding novel techniques such as genome sequencing \[18, 20\]).

It is well known that the genome codes for proteins in a redundant way. With the exception of methionine and tryptophan, different codons codify for the same amino acid, which implies that not one, but an ensemble of nucleotide sequences, after translation, give the same protein. This is not, however, the only source of redundancy during the translation of a genomic sequence into a phenotype. Redundancy in translation can also be a consequence of the existence of a synonymous secondary structure (the neutrality upon a change of amino acid) \[21\] or the duplicity of entire genes \[22\], among other factors \[23\]. There are several ways in which a deficiency produced by the lack of a given protein (which can happen when a complete gene is deleted from the genome of an individual) can be overcome by using alternative regulatory pathways or regulatory networks \[22, 24\]. These mechanisms are the subject of study of so-called genetic robustness \[23, 25–28\]. All the observed mechanisms for robustness based on redundancy support the hypothesis that the genome of a species is not a singular sequence, but a process producing an ensemble of admissible nucleotide sequences.

We will understand the structure of the genome as the set of rules characterizing the ensemble of nucleotide sequences allowed in a specific species. In order to make the analysis easier, we will assume that the sequences contained in such an ensemble are infinite and that the ensemble itself is translationally invariant. Hence, we can see the ensemble of all the allowed nucleotide sequences of a given species as a shift space \[29\]. A shift space consists of a set of infinite or semi-infinite sequences, with values in a finite alphabet, which is closed under the action of the left shift (the shift mapping). A shift space can be characterized in several different ways, for example, by means of its language, i.e., the set of words occurring as subblocks of allowed sequences (see figure 1), or by means of a set of forbidden blocks, which is a set of subblocks which does not appear in any sequence of the shift space. The theory of shift spaces seems to be suitable for understanding DNA structure. In particular, we will be concerned with the complexity function of symbolic systems, i.e., the symbolic complexity (SC). The symbolic complexity of a shift space is the function which associates the number of words of length \(\ell\) (called \(\ell\)-words hereafter) occurring as subblocks of the allowed sequences with each integer \(\ell\).

\[ x = \text{symbolic sequence} \]

\[ a = \text{word or subblock} \]

\[ \text{Figure 1. A schematic representation of words or subblocks occurring in a symbolic sequence.} \]
The diversity of \( \ell \)-words gives information about the structure of the symbolic system, or, in other words, the mechanisms that generate such a string. The problem of determining the SC from a finite (nucleotide) sequence has been considered in [30, 31]. As explained in such references, for a finite sequence, the complexity function has a well-defined behavior: it is an increasing function for small word lengths and it becomes a decreasing function for large word lengths. However, such a complexity function is not representative of the complexity function of the process, since, as it is known, the complexity function is always increasing [29]. In this paper, our approach intends to give a better approximation of the complexity function of the process, for which the standard way of estimating the complexity function—as done in [30, 31]—is not completely suitable, especially for large word lengths.

It is known that the standard way of estimating the SC requires a large sample (large sequence) to have a good estimation. As we show below, we can overcome this difficulty if we consider the SC of the process, instead of the SC in the strict sense. The purpose of this work is twofold: first we introduce an estimator for the symbolic complexity of the process, which gives accurate estimations even with a relatively small sample; then, we use this estimator to study the SC of real genomes, particularly that of the Hominidae family. We found that the SC behaves in a characteristic way along all the genomes we considered: it consists of an exponential part for words of short length, followed by a linear part for words of longer lengths. This means that the diversity of words up to a given size is very large; in particular, all the words up to size 10 are present in the genome. The linear part reflects the relatively low diversity of large words, which is a sign of the presence of highly ordered long-range structures in the genome. Moreover, our findings suggest that the SC is characteristic of each species, in which case this property would be a significant indicator in the study of the molecular evolution of a species. In order to interpret the form of the SC observed in real genomes, we analyze the SC of sequences generated by a simple evolution model. This allows us to relate the characteristic form of the complexity function of real genomes to the possible mechanisms underlying genome evolution.

The paper is organized as follows. In section 2 we present a new estimator for SC which we test over several symbolic systems. In section 3 we use our estimator to obtain the SC of the coding DNA sequences of several species of the Hominidae family. In section 4 we estimate the SCs for ensembles of sequences obtained from a DNA evolution model at finite time and compare them with the SCs obtained from real DNA sequences. We close the paper in section 5 with a brief discussion and concluding remarks.

2. The complexity function of the process

The point of view that we adopt here is as follows. First we assume that the genome is a stationary process over the finite set \( \mathcal{A} = \{A, G, T, C\} \), described by the shift-invariant measure \( \mathbb{P} \). Thus, we should consider the complexity function as a random variable. Actually, in real genomes, the complexity function in the strict sense would have no meaning, since this quantity may be highly variable from one individual to another. Hence, the expected value of the diversity of words would better be able to characterize this property for a species. Then, we interpret the genome of a species as a result of a stochastic process that can be characterized by means of its typical\(^3\) symbolic complexity \( C(\ell) \). Hence, the mean value \( C(\ell) \) of the SC should be estimated from an ensemble of realizations of the process. The

\(^3\) Typical with respect to the corresponding invariant measure \( \mathbb{P} \).
corresponding estimator, to be defined below, will be obtained by considering the number of different ℓ-words occurring in a sample of size \( m \).

### 2.1. An estimator for the symbolic complexity

As mentioned before, the genome is interpreted as a result of a stationary process and is described by the shift-invariant measure \( \mathbb{P} \). The support of \( \mathbb{P} \) is a symbolic system \( \mathcal{X} \) characterized by its language \( \mathcal{L}(X) \), i.e., the set of all words occurring as subblocks of admissible sequences on \( X \). If we denote by \( \mathcal{L}_\ell(X) \) the set of all \( \ell \)-words occurring as subblocks of admissible sequences on \( X \), then the SC of \( X \) is the function which associates the cardinality of \( \mathcal{L}_\ell(X) \) with each \( \ell \in \mathbb{N} \). Within this framework, the genome of an individual can be considered as the observation of a finite random sample with finite precision. The random sample could be composed of a single sequence. In any case, we assume that the sequences in the sample are typical with respect to \( \mathbb{P} \), which in turn is assumed to be ergodic with respect to the shift map. Besides ergodicity, a shift-invariant measure \( \mathbb{P} \) describing nucleotide production is required to satisfy an equidistribution property to be specified below.

The problem we face can be stated as follows: given a sample \( S \) of not necessarily different \( \ell \)-words, estimate the complexity \( C(\ell) \). For this, let us consider the random variable \( N_{\ell,s} \) to be defined as the number of different \( \ell \)-words occurring in a random sample \( S \) of size \( s \).

We require the process to satisfy the following equidistribution property: all possible \( s \)-sets of admissible \( \ell \)-words have the same probability of occurring\(^4\). Assuming this principle, it is shown in appendix A that

\[
\mathbb{P}(N_{\ell,s} = n) = \binom{s - 1}{n - 1} \binom{C(\ell)}{n} \binom{C(\ell) + s - 1}{s}.
\]

The expected value for \( N_{\ell,s} \) is readily computed, giving

\[
\mathbb{E}[N_{\ell,s}] = \frac{C(\ell)s}{C(\ell) + s - 1}.
\]

The above equation clearly shows that \( \mathbb{E}[N_{\ell,s}] \to C(\ell) \) when \( s \to \infty \). On the other hand, a direct count of different \( \ell \)-words in \( S \) is not a good estimator for the complexity function. Indeed we can observe that \( N_{\ell,s} \) is not an unbiased estimator for \( C(\ell) \) since \( \mathbb{E}[N_{\ell,s}] < C(\ell) \) for all sample sizes \( s \in \mathbb{N} \).

The variance of \( N_{\ell,s} \) can also be computed, giving

\[
\text{Var}[N_{\ell,s}] = \frac{C(\ell)(C(\ell) - 1)s(s - 1)}{(C(\ell) + s - 1)^2(C(\ell) + s - 2)}.
\]

Hence, the variance of \( N_{\ell,s} \) is small whenever \( C(\ell) \gg s \) and actually, in this regime

\[
\frac{\sqrt{\text{Var}[N_{s,s}]} \text{Var}[N_{s,s}]}{\mathbb{E}[N_{s,s}]} = \mathcal{O}\left( \frac{1}{\sqrt{C(\ell)}} \right).
\]

This means that most realizations of \( N_{\ell,s} \) result in a value which does not deviate significantly from \( \mathbb{E}[N_{\ell,s}] \) due to ‘random fluctuations’. Hence, even in the case where the typical complexity \( C(\ell) \) is larger than the sample size \( s \), the random variable \( N_{\ell,s} \) would give us

\[\text{\footnotesize \footnote{Actually, this hypothesis can be weakened by requiring an approximated equidistribution instead of a strict equidistribution in the sense established above.}}\]
accurate information about $C(\ell)$, that we can extract from $\mathbb{E}[N_{c,s}]$ because of the small random fluctuations. Inverting equation (2) we obtain

$$C(\ell) = \frac{(s - 1)\mathbb{E}[N_{c,s}]}{s - \mathbb{E}[N_{c,s}]}$$

and the above reasoning gives us a basis to propose

$$c_{c,s} = \frac{s N_{c,s}}{s + 1 - N_{c,s}}$$

as a statistical estimator for the typical SC $C(\ell)$. Here we have used $N_{c,s}$ as an estimator of its mean, and we have replaced $s$ by $s + 1$ in order to avoid divergences. In appendix A we show that $c_{c,s}$ has the expected value given by

$$\mathbb{E}[c_{c,s}] = C(\ell) - (C(\ell) - s)\mathbb{P}(N_{c,s} = s).$$

Note that $c_{c,s}$ is unbiased for $s \geq C(\ell)$ and that $\mathbb{E}[c_{c,s}]$ approaches $C(\ell)$ when the probability of having no repeated word in the sample is negligible. Note also that for samples of large cardinality with respect to $C(\ell)$, we have $c_{c,s} = C(\ell) + O(C(\ell)^2/s)$. Therefore $c_{c,s}$ is a good estimator of the SC for processes whenever the sample size is large enough compared to the actual SC, or for processes with high complexity satisfying the equidistribution properties describe above.

### 2.2. Examples

In order to test this estimator in samples coming from systems with fast complexity growth, we have to ensure that we have the conditions required for the validity of equation (1), namely, that all possible $s$-sets of admissible $\ell$-words have the same probability of occurring. This condition is approximately fulfilled for processes with maximal entropy on symbolic spaces satisfying the specification property. Indeed, let us suppose that $X$ is a specified subshift and that the process describe by $\mathbb{P}$ is the unique one maximizing the entropy. Let $x = x_1 x_2 \cdots \in X$ be a typical sequence for $\mathbb{P}$, and consider the sample set

$$S = \{x_{(t+1)}, x_{(t+2)}, \cdots, x_{(t+\Delta)}\} \subset \mathcal{L}_{\ell}(X).$$

It is well known that under the conditions stated above, $\mathbb{P}$ satisfies the following asymptotic independence,

$$\mathbb{P}(S) \approx \prod_{k=0}^{\ell-1} \mathbb{P}(x_{(t+\Delta)+k}),$$

whenever the gap $\Delta$ is sufficiently large. On the other hand, maximal entropy implies an approximative equidistribution, i.e., for all $\ell$ and $x_1 x_2 \cdots x_{\ell}, y_1 y_2 \cdots y_{\ell} \in \mathcal{L}_{\ell}(X)$, we have

$$\frac{\mathbb{P}(x_1 x_2 \cdots x_{\ell})}{\mathbb{P}(y_1 y_2 \cdots y_{\ell})} = O(1),$$

hence, by taking a sufficiently large $\Delta$ with respect to $s$, we obtain

$$\mathbb{P}(S) = O(C(\ell)^{-s}).$$

Thus, the sample $S$ described above can be used to estimate the typical complexity $C(\ell)$, whenever the $\ell$-word separation $\Delta$ is sufficiently large. In order to test our estimator we

---

5 These concepts are clearly explained in [32].

6 It should suffice to take $\Delta = O(\log(s))$. 

---
implemented this sampling procedure on typical sequences of well-characterized binary processes. First, we produced long sequences of symbols from three different systems: the full shift (the Bernoulli scheme on two symbols), the Fibonacci shift (sequences with the forbidden word 00), and the run-limited length shift (as an aperiodic shift, with a countable infinite set of forbidden words [29]). In every case the sequences were produced at random according to the corresponding probability measure of maximal entropy. For each one of these three typical sequences, we have taken a sample of words of lengths ranging from 1 to 50. The fitting \( c_{\ell s} \approx \exp(h \ell) \) allows us to estimate the entropy of the process. The estimated values are \( h = 0.721 \pm 0.0025 \) for the full shift, \( h = 0.461 \pm 0.0014 \) for the Fibonacci shift, and \( h = 0.384 \pm 0.0012 \) for the \((1, 3)\) run-length limited shift. These estimations are in good agreement with the exact values of the corresponding entropies.

In figure 2, we show the values of \( N_{\ell s} \) and \( c_{\ell s} \) as functions of \( \ell \), obtained by using the sampling method described above. Note that the function \( \ell \mapsto N_{\ell s} \) exhibits a ‘kink’ (an abrupt change in the monotony), a behavior that has been previously observed in [31] (which would also be expected from the results by [30, 33–35]). This is consistent with the behavior predicted by equation (2), which can be calculated for these cases since we know the exact value of \( C(\ell) \). From \( N_{\ell s} \), we obtain the values of \( c_{\ell s} \) which, as mentioned before, is a good estimate of \( C(\ell) \). It is known that \( C(\ell) \) grows exponentially with \( \ell \) in each one of the analyzed examples. Hence, from \( c_{\ell s} \) we can estimate the entropy \( h \) of the process by fitting \( c_{\ell s} \approx \exp(h \ell) \). The entropy of the three analyzed examples can be exactly computed, giving \( h = \log(2) \approx 0.69314 \) for the full shift, \( h = \log((1 + \sqrt{5})/2) \approx 0.48121 \) for the Fibonacci shift, and

\[
h = \log(\sqrt{29 + \sqrt{837}}/52 + \sqrt{(29 - \sqrt{837})/52 + 1/3}) \approx 0.39197
\]

for the \((1, 3)\)-run-left limited shift. From the curves of \( c_{\ell s} \) shown in figure 2, we obtained the corresponding estimations for the topological entropies by means of the least squares method:

![Figure 2. Estimations of the complexity functions for three symbolic systems: the (1, 3)-run-length limited (solid lines), the Fibonacci (dashed lines) and the full shift (dot-dashed lines). For each one of these systems, the random variables \( c_{\ell s} \) (red lines) and \( N_{\ell s} \) (black lines) were computed from a typical sequence of maximal entropy of length \( 6 \times 10^6 \), from which we extracted samples of \( s = 10^5 \) sub-words of lengths ranging from 1 to 50. The fitting \( c_{\ell s} \approx \exp(h \ell) \) allows us to estimate the entropy of the process. The estimated values are \( h = 0.721 \pm 0.0025 \) for the full shift, \( h = 0.461 \pm 0.0014 \) for the Fibonacci shift, and \( h = 0.384 \pm 0.0012 \) for the (1, 3) run-length limited shift. These estimations are in good agreement with the exact values of the corresponding entropies.](image-url)
for the full shift, \( h = 0.461 \pm 0.0014 \) for the Fibonacci shift, and \( h = 0.384 \pm 0.0012 \) for the (1, 3)-run-length limited shift. Note that the better estimation corresponds to the one for which the topological entropy is the lowest, which is due to the fact that for sequences of larger entropy, the value of \( N_{l,s} \) ‘saturates’ faster. Saturation occurs when the expected value of \( N_{l,s} \) differs by less than one from the sample size \( s \).

3. Symbolic complexity for coding nucleotide sequences

We use the method described above to estimate the SC of coding DNA sequences. Coding DNA has some similarities with shift spaces in the sense that, for example, coding DNA does not necessarily admit all the possible words we can build with A, T, C and G. This is because of the existence of deleterious mutations, implying that some ‘words’ on the genome are not allowed for living organisms. Thus, we think that DNA could be characterized by some language through a set of forbidden words. Then, the SC function might give some information on such a structure.

In figure 3, we display the SC for word lengths \( l \) in the range \( 1 \leq l \leq 50 \) bp, which was obtained from coding DNA corresponding to Pan troglodytes, Gorilla gorilla gorilla, Pongo abelii and Macaca mulatta. The sample used to estimate the SC is \( s = 10^5 \) and this sample was taken from the coding region of chromosome 1 of each species. Such coding DNA sequences were taken from the GenBank database [37]. Each sequence obtained from the database had a size of \( 6 \times 10^6 \) bp.

Figure 3. Estimates of the symbolic complexity for species belonging to the Hominidae family. The SC was estimated from a nucleotide sequence \( 6 \times 10^6 \) bp long taken from the first and second chromosomes (whenever necessary to complete the mentioned length) corresponding to coding regions. We took a sample of \( s = 10^5 \) \( l \)-words with \( 1 \leq l \leq 50 \). We can appreciate how the SC corresponding to human DNA is lower than the rest of the Hominidae. Indeed, the order observed in the complexity correlates positively with a phylogenetic distance estimated from other means (see, for example, [36]). The inset shows the exponential growth in complexity, common to all the analyzed sequences, for word lengths up to a size of \( l \approx 10 \).

\( \hat{h} = 0.721 \pm 0.0025 \) for the full shift, \( \hat{h} = 0.461 \pm 0.0014 \) for the Fibonacci shift, and \( \hat{h} = 0.384 \pm 0.0012 \) for the (1, 3)-run-length limited shift. Note that the better estimation corresponds to the one for which the topological entropy is the lowest, which is due to the fact that for sequences of larger entropy, the value of \( N_{l,s} \) ‘saturates’ faster. Saturation occurs when the expected value of \( N_{l,s} \) differs by less than one from the sample size \( s \).

7 It is necessary to notice that the confidence bounds correspond to those obtained from the linear fit. These confidence bounds do not take into account the variance due to the realization of the process. Improved confidence bounds can be obtained by estimating the SC for several realizations of the string.
Figure 4. Estimates of the SC for coding DNA for several values of the sample size. The SC was estimated from a nucleotide sequence $6 \times 10^6$ bp long taken from chromosome 1. We took a sample of $s = 10^5$ $\ell$-words with $1 \leq \ell \leq 50$. We used sample sizes of $5 \times 10^4$, $6 \times 10^4$, $7 \times 10^4$, $8 \times 10^4$, $9 \times 10^4$, and $10^5$. We can appreciate how the SC diminishes as the sample size increases, with the exception of the SC for $s = 10^5$. This would imply that the SC seems to converge to a fixed complexity function as the sample size increases.

Figure 5. The symbolic complexity for the genome of *Homo sapiens* (black lines) and *Pan troglodytes* (red lines). Each curve corresponds with an estimation of the SC by means of the estimator given in equation (4). For each curve we used one, two or more chromosomes in order to complete a sample string $6 \times 10^6$ bp long. From such a string we took a sample of $10^5$ words of $\ell$ bp for every $1 \leq \ell \leq 50$. The inset shows that all the chromosomes have complexity functions with exponential growth at small word sizes.
We calculated the corresponding values of $c_{\ell,s}$ for every $\ell$ in the selected range. In the figure, we appreciate how the human coding sequences have the lowest complexity amongst the analyzed species. We should also notice the progressive increase in complexity as the species moves away from human—in the phylogenetic sense—according to the reported phylogenetic trees [36]. We can also appreciate a pattern in the increase in complexity present in all the analyzed organisms. For small word lengths ($\ell \leq 10$), all possible ‘genomic words’ appear. This is clear from the exponential growth $c_{\ell,s} \approx 3.94^\ell$ with a correlation coefficient 0.99. On the other hand, for large enough word lengths (12 $\leq \ell \leq 50$), the symbolic complexity grows linearly.

In figure 4, we plot the SC for the coding region of chromosome 1 of Homo sapiens for several values of the sample size $s$. We use the following values for $s$: $5 \times 10^4$, $6 \times 10^4$, $7 \times 10^4$, $8 \times 10^4$, $9 \times 10^4$, and $10^5$. From the plotted SCs we can observe that it decreases as the the sample size increases, with the exception of the last value, $s = 10^5$. This would mean that the SC seems to converge to a definite value. Unfortunately, due to the definiteness of the sample, we cannot obtain a reliable SC for larger sample sizes than $s = 10^5$.

Now we proceed to observe how the SC behaves along all the coding regions of the whole genome (see figure 5), to appreciate the complexity functions obtained from the coding regions of almost all the chromosomes of Homo sapiens and Pan troglodytes. We found that the estimated SCs seem to have the same qualitative form in all the genome species studied here. Moreover, from this analysis we observe how, on average, the complexity functions for both species differ. This would mean that the complexity function, in the sense of the diversity of $\ell$-words present in the genome, seems to be characteristic for each species. The differences of SC between different species could mainly be associated with the point mutation processes. Indeed, we would expect that the higher the mutation rate, the higher the diversity of words that will be present in the genome.

The linear growth in complexity suggests that the genomic sequences are highly ordered, which is as it occurs for almost periodic symbolic sequences. Symbolic systems having linear complexity have zero topological entropy, which is a characteristic of topologically predictable dynamics [38]. These kinds of symbolic sequences might be generated by a deterministic process in the sense of topological dynamics (see, for example, [38] and the references therein). In the literature, it can be found that several symbolic dynamical systems have this characteristic, among which we find the Thue–Morse, the Toeplitz and the Cantor shifts [39, 40].

The above-mentioned symbolic systems are the result of a substitutive process. It is known that any substitution process leads to symbolic dynamical systems having a linear SC and, therefore, zero topological entropy (see, for example, proposition 5.12, page 148 in [41]). Our analysis of coding DNA sequences suggests that these nucleotide sequences could be the result of a substitutive process. Indeed, several authors have proposed random substitution systems to model DNA evolution [8–10, 13].

4. Symbolic complexity for some evolution models

In this section we briefly review some models of genome evolution. For each case we present the corresponding estimation for the SC as a function of the word length.

The expansion-modification system. One of the earliest models of genome evolution is the well-known expansion-modification system, originally proposed by Li [10]. This system was proposed in order to understand the origin of the long-range correlations observed in real genomic sequences [3, 4, 42]. Recently, the expansion-modification system has been studied
rigorously within the formalism of symbolic dynamics [43]. In such work it has been proved that such a system has a unique stationary measure that is reached by any initial condition. It has also been found that the correlation function has a polynomial decay for almost all the parameter values. Moreover, the unique stationary measure is supported on the full shift, which means that all the ℓ-words are typically present in any realization of a sequence produced by the expansion-modification system. This implies that the SC for the expansion-modification system equals the one for random sequences for all the values of the expansion probability \( p \). This property is not compatible with the observed complexity from real genomic data.

To obtain an estimation of the SC we simulate the expansion-modification process defined as follows. First, we give a seed, i.e. a symbol taken from the binary alphabet \( \mathcal{A} = \{0, 1\} \). Then, the seed is duplicated (or ‘expanded’) with probability \( p \) and mutated (or ‘modified’) with the complementary probability \( 1 - p \). We repeat this process successively with all the symbols present in the string. We stop the process until the sequence has reached a size of \( 10^6 \) symbols. After that we obtain the SC following the steps described in section 2. In figure 6, we plot the complexity function for the expansion-modification system for several values of the expansion probability. We can appreciate that the complexity function behaves exponentially, and a better estimation is obtained when the expansion probability \( p \) equals \( 1/2 \). Actually, the SC for the expansion-modification system should be same for all the values of \( p \) as stated above. Therefore, we have that the random variable \( K \) defined above is not a good estimator for the complexity function. The latter is a consequence of the dependence between words in the sample, a phenomenon that clearly occurs due to the long-range correlations. Moreover, since the system takes a long time to reach its stationary state [43], it is clear that not all the words would be present in the realization of a symbolic sequence. The latter could also affect the estimation of the complexity, even if the estimator is accurate. In any case, the estimated complexities from the sequences obtained by the expansion-modification system do not follow the profile we have observed in real genomic sequences. However, we should
emphasize that the higher the duplication probability, the lower the SC we observe in the sequences.

The Zaks model. A genome evolution model that reproduces the three-base periodicity was put forward by Zaks [13]. The Zaks model can be seen as a substitutive system that is randomly perturbed. The base substitution system has a fixed point that turns out to be a Thue–Morse sequence, a system that has a linear complexity as mentioned above. The perturbations introduced in this system have the effect of randomizing the Thue–Morse sequence, which consequently modifies the original complexity function. This results in a sequence having a correlation function with singular and continuous components. These are intended to model the three-base periodicity and the long-range correlations respectively, both present in real genomes.

We simulated the Zaks model for several values of the involved parameter. Specifically, we simulated the process that produces symbolic sequences by means of the following (stochastic) algorithm. Given a ‘seed’ \( x \) (a symbol belonging to \( \mathcal{A} = \{0, 1\} \)) we have that,

\[
x \rightarrow \begin{cases} 
xx \text{ with probability } p, \\
x1 \text{ with probability } 1 - p,
\end{cases}
\]
where $\mathcal{F}$ stands for the complementary symbol, i.e., $\mathcal{F} = 0$ and $\mathcal{F} = 1$. This process is further extended coordinate-wise to finite or infinite chains. It is worth noticing that for $p = 0$ the process results in a Thue–Morse sequence $[13, 41]$. In contrast, for $p = 1$ the resulting symbolic string becomes $x = \ldots 0$, with probability 1, in the limit of infinite iterations of the process. In figure 7 we observe that the SC again does not reproduce the profile for the complexity function observed in real genomic data. Actually, for a small duplication probability $p$ the complexity function fits better to a power law, i.e., $C(l) \propto l^{\alpha(p)}$ for word lengths larger than 10. From our simulations, we obtain that $\alpha(0.01) \approx 2.33 \pm 0.012$, $\alpha(0.02) \approx 2.26 \pm 0.016$, $\alpha(0.03) \approx 2.92 \pm 0.02$, and $\alpha(0.04) \approx 3.19 \pm 0.018$. For larger duplication probabilities $p$, the complexity function fits better with an exponential, $C(l) \propto \exp(\beta(p)l)$. We obtain the following values from the simulations: $\beta(0.1) \approx 0.16 \pm 0.001$, $\beta(0.2) \approx 0.358 \pm 0.003$, $\beta(0.3) \approx 0.546 \pm 0.003$, and $\beta(0.1) \approx 0.683 \pm 0.004$. In the latter case, we can observe that the complexity function profile is similar to the one obtained for the expansion-modification system. Again, we observe that the duplication process has the effect of decreasing the complexity function.

The Hsieh model. The evolution model proposed by Hsieh et al in [8] consists of two random processes: single base replacement (i.e., a single base mutation) and a random segmental duplication. These processes occur naturally in real genomes (see, for example, [44, 45] and the references therein) and allow the sequences to grow exponentially fast with random mutations. The length of the segments to be duplicated is not constant. Thus, in every
step of the evolution a random length is chosen from a prescribed exponential distribution function. The authors of [8] introduced such a model in order to understand the origin of the $k$-mer distributions that they observed in real nucleotide sequences. We obtain symbolic sequences from this model as follows. First, we produce a random sequence $L_0 = 1000$ symbols long; the symbols are taken from the alphabet $\mathcal{A} = \{A, T, C, G\}$. Then, the sequence is subjected to the following evolution processes: i) single base replacements (SBR) and ii) segmental random duplications (RD). As in [8] we define $\eta$ as the ratio of chances of having an SBR event with respect to an RD process. We evolve the symbolic sequence in discrete time. At each step the sequence undergoes either an SBR or an RD. If $\eta > 1$ the probability of an RD occurring is $1/\eta$. If an SBR occurs we choose a random site $k_m$ with uniform distribution in $\{1, 2, \ldots, L_c\}$. The symbol at $k_m$ is replaced by one of the three complementary symbols with probability $1/3$. If an RD occurs we choose at random a duplication length $L$ from a probability distribution $p(L) \propto \exp(-\sigma L/L_c)$. Here, $L_c$ is the current length of the symbolic string and $\sigma$ a parameter playing the role of a length scale for the duplicated segments. Then, we choose a random site $k_c$ with uniform distribution in $\{1, 2, \ldots, L_c - L\}$ and copy the segment from the site $k_c + 1$ to the site $k_c + L$. Next, a second random site $k_d$ is chosen with uniform distribution in $\{1, 2, \ldots, L_c\}$, where the copied segment is inserted. To perform our simulations we choose the same parameter values as those used by Hsieh et al in [8] for their simulations.

In figure 8(a) we plot the estimated SC for two values of the mutation-duplication ratio $\eta = 500$ and $\eta = 600$ both with $\sigma = 15000$ fixed. In this case, we can observe that the profile of the complexity function is qualitatively similar to the one for coding genomic sequences. The complexity function for these parameter values starts with an exponential growth (for word lengths in the range $1-5$) eventually becoming linear (for word lengths in the range $7-50$). In figures 8(b) and (c) we plot the SC for the system without SBR events ($\eta = 0$) and short duplication segments ($\eta = 500$ and $\sigma = 25$). We should notice that despite the absence of SBR we obtain a large diversity of $\ell$-words (i.e., a large SC). This observation lets us conclude that the SBR is not the sole source of SC, indeed we can appreciate that random duplication events of short polynucleotides can also give rise to a huge diversity of $\ell$-words.

The Messer model. The model of Messer et al [11] goes a step further by considering not only random mutations and segmental duplications, but also single base deletions and random insertions. Indeed, they are able to obtain an explicit expression for the correlation function of such a model, obtaining polynomial decay as expected. The main difference with the model introduced by Hsieh et al is that the segmental random duplications have a fixed length and that the system evolves continuously in time. In this work, we simulate the Messer model in discrete time since we are only interested in the final state (a string $10^6$ symbols long) and not how such a state has been reached. The procedure is as follows. First we produce a random string $L = 1000$ symbols long. Then, the sequence is subjected to the following random processes: (i) mutations, (ii) duplications, (iii) insertions and (iv) deletions. The random mutations and random duplications correspond to the SBR and RD events implemented in the Hsieh model. For the random duplications, the Messer model uses a fixed length for the duplicated segments which we call $L_d$. The insertions are carried out as follows: first we chose at random a site $k_i$ in $\{1, 2, \ldots, L_c\}$ (where $L_c$ is the current length of the string) with the uniform distribution. Then at the site $k_i$ a random word of length $L_i$ is inserted. The deletions occur in a similar way: we choose at random a site $k_d$ in $\{1, 2, \ldots, L_c\}$ with uniform distribution, and all the letters in the sites $k_d$, $k_d + 1$, $\ldots$, and $k_d + L_d - 1$ are suppressed. Here, the length $L_d$ of the suppressed word is constant. In the original model put forward by Messer in [11] the mutation, duplications, insertions and deletions occur at constant rates represented
by $\mu$, $\delta$, $\gamma^+$ and $\gamma^-$ respectively. In our simulations we evolve the string as follows: at every step the string undergoes one of the processes (i), (ii), (iii) or (iv) with a given probability. The sum of the probabilities corresponding to the four processes equals one. The relation between the rates and the probabilities is as follows:

First let $mdg_g = \mu + \delta + \gamma^+ + \gamma^-$ be the sum of all the rates. The probability of having a random mutation $p_0$ is given by $p_0 = \mu/s$, the probability of having a random duplication is $p_1 = \delta/s$, the probability of having a random insertion is $p_2 = \gamma^+/s$ and the probability of having a random deletion is $p_3 = \gamma^-/s$.

For our simulations, we choose the rates $\mu = 1$, $\delta = 25$, $\gamma^+ = 5$ and $\gamma^- = 10$, which correspond to one of the numerical experiments performed in [11]. Then, we only vary the length of the duplications, insertions and deletions.

In figure 9(a) we show the estimated SC for the Messer evolution model for two different sets of parameter values {\( L_d = 1000, L_i = 10, L_s = 10 \)} and {\( L_d = 1000, L_i = 50, L_s = 40 \)}. In this figure, we can appreciate that the complexity function has a profile similar to the one found in the Hsieh model. The behavior of the complexity in both cases resembles the one found in real genomic data from coding sequences. In figure 9(b) we show the estimated SC for the sets of parameter values {\( L_d = 50, L_i = 20, L_s = 10 \)} and {\( L_d = 100, L_i = 10, L_s = 10 \)}. In this case, we observe again the same phenomenon found for the Hsieh model: the shorter the duplication segments, the larger the SC.

The Massip–Arndt model. The evolution model which seems to better reproduce the characteristics found in real genomes is the one introduced by Massip and Arndt in [46]. This model incorporates two basic mechanisms: point mutations and duplications. These mechanisms are incorporated as follows [46]. First we take a random sequence $x = x_1x_2...x_n$ of $n$ symbols long, each symbol $x_i$ taken from the alphabet $A = \{A, T, C, G\}$ representing the set of nitrogenous bases of DNA. Then, such a sequence is evolved in discrete time by the following stochastic algorithm. At each step, the sequence undergoes a single base replacement (SBR) with probability $p_{SBR}$ or a segmental duplication (SD) with probability $p_{SD}$. The SBR process is achieved by taking at random a basis (a symbol) along $x$ and replacing it with one of the remaining bases, all of which have an equal probability of being chosen.
Figure 10. The complexity function for the Massip–Ardnt genome evolution model. Each curve corresponds to an estimation of the complexity function by means of the estimator given in equation (4) for different values of SBR probability. We used the parameter values $p_{\text{SBR}} = 0.050$ (solid line), $p_{\text{SBR}} = 0.010$ (dashed line), and $p_{\text{SBR}} = 0.001$ (dot-dashed line). The length of the duplicated segment is fixed to $L = 10^{3}$ and the SD probability is fixed to $p_{\text{SD}} = 0.600$. We notice that the larger the mutation probability, the larger the complexity we obtain. All the profiles reproduce the exponential-linear behavior of the complexity function.

Figure 11. The complexity function for the Massip–Ardnt genome evolution model. Each curve corresponds to an estimation of the complexity function by means of the estimator given in equation (4), for different values of the SBR probability. We used the parameter values $p_{\text{SBR}} = 0.050$ (solid line), $p_{\text{SBR}} = 0.010$ (dashed line), and $p_{\text{SBR}} = 0.001$ (dot-dashed line). The length of the duplicated segment is fixed to $L = 10^{3}$ and the SD probability is fixed to $p_{\text{SD}} = 0.600$. We notice that the larger the mutation probability, the larger the complexity we obtain. All the profiles reproduce the exponential-linear behavior of the complexity function.
Specifically, the process is as follows. We take a random number \( j \) from the set \( \{1, 2, \ldots, n\} \) with uniform probability. Then, we replace the symbol \( x_j \) with one of the three complementary symbols in the alphabet, each symbol having a probability of \( \frac{1}{3} \) of being chosen. The SD process consists of duplicating an entire segment, of length \( L \), occurring in the sequence \( x \). To this end, we first choose independently two random numbers \( j, k \) from the set \( \{1, 2, \ldots, n\} \) with uniform probability. Then, the segment \( x_j x_{j+1} \ldots x_{j+L-1} \) is replaced by a copy of the segment \( x_k x_{k+1} \ldots x_{k+L-1} \). In this way, after the SD process there is an entire copy of the segment \( x_j x_{j+1} \ldots x_{j+L-1} \), if the chosen segments do not overlap.

We simulate the SBR and SD processes to produce sequences from which we estimate the symbolic complexity. The parameters we use are the following. We fix the length of the sequence \( x = x_1 x_2 \ldots x_n \) to \( n = 6 \times 10^6 \) and the duplicated segment length to \( L = 1000 \). The initial sequence is chosen randomly with all the bases \( \mathcal{A} = \{A, T, C, G\} \) independently and having equal probability of occurring. Then, the random sequence \( x = x_1 x_2 \ldots x_n \) is subject to the SBR and SD processes with the following set of parameter values: \( p_{\text{SBR}} = 0.050, p_{\text{SD}} = 0.600 \), \( p_{\text{SBR}} = 0.010, p_{\text{SD}} = 0.600 \), and \( p_{\text{SBR}} = 0.001, p_{\text{SD}} = 0.600 \). We performed a total of \( 10^4 \) time steps.

In figure 10 we observe how, for this model, the estimated complexity behaves in a similar way to the one we found in the real genome. Moreover, the numerical experiments we performed clearly reveal that the larger the SBR probability, the larger the SC. This would indicate that in real genomes, species with a lower SC might be subject to lower mutation rates.

Next, to understand the turnover from exponential to linear behavior in the SC curve, in figure 11 we display the curve for the SC for several values of \( L \), the length of the duplicated segment in the Massip–Ardnt model. For all the SC curves we used the same time-steps for the simulations. From this figure, it is clear that the shorter the duplication segment, the larger the diversity of \( \ell \)-words. Let us call \( \ell_c \) the crossover from exponential to linear behavior. We should notice that for \( L = 10^2 \) the crossover occurs approximately at \( \ell = 12 \) for \( L = 10^3 \) at
Finally, for the Massip–Ardnt genome evolution model we show how the SC evolves in time. In figure 12, we plot the SC for several values of the time step in the evolutionary process. First, we evolve a random string 10^6 bp long with the Massip–Ardnt process. From this process we obtain the symbolic sequences at $t = 2 \times 10^3$, $t = 5 \times 10^3$ and at $t = 9 \times 10^3$. Then, we estimate the SC by extracting a sample of $10^5$ from such a sequence. We can observe that the complexity decreases as $t$ increases. This result shows that the process has not reached the stationary state. This is clear from the fact that the curve depends on time. Moreover, to our knowledge, we do not know if the Massip–Ardnt process actually reaches a stationary state. Nevertheless, we can appreciate that the SC curve seems to preserve the exponential-linear behavior for a wide interval of time steps. It is clear, of course, that a more detailed study on the evolution in time of the SC is necessary for the Massip–Ardnt evolution model. This would give a deeper insight into which processes are relevant in the evolution in time of real genomes.

5. Conclusions

We have proposed a way of estimating the complexity function of symbolic systems from finite strings, which we used to analyze the coding DNA from the genomes of four species belonging to the Hominidae family. This study gave us information about the structure of the complexity function of the genome, which shows a characteristic shape common to all the species analyzed here. We found that the complexity function follows an exponential growth for words in the range 1–10 bp, followed by linear growth for words in the range 11–50 bp. The latter is a kind of behavior associated with a process generating long-range order in the system, such as it occurs in substitutive dynamical systems [39, 40]. This ordering derives from the segment duplication process in the evolution model that we analyzed. We also observed that the complexity function for real genomes is specific for a given species. The differences in SC between different species could be associated with their evolutive history—particularly with the difference in the accumulation of point mutation processes. This interpretation is supported by the observations made in the evolution model we studied, where an increase in the diversity of observed words is associated with an increase in the probability of occurrence of a single base replacement. In other words, the higher the mutation rate, the larger the SC we observe. This fact can easily be understood since the mutation rate has the effect of producing new words that were not originally present in the sequence.

In summary, the estimation of the complexity function for real genomes could shed light on the nature of the evolutive mechanisms that led the genome to its current structure. The analysis of a particular evolution model gives some clues concerning how evolution history shapes the profile of the complexity function. The combined study of the SC of different evolution models and real genomes would allow us to unravel some of the key mechanisms governing the evolution of the genome. This will require a more extended and detailed study, of which the present represents a first step.

Acknowledgments

This work was supported by CONACyT through grants CB-2012-01-183358 and CB 2014-237324-F. R S-G thanks F Vázquez for carefully reading the manuscript and for giving useful comments on this work.
Appendix A. Some properties of the random variables $N_{\ell,s}$ and $c_{\ell,s}$

Let $C(\ell)$ be the typical SC of $X \subset A$ with respect to the shift invariant ergodic measure $\mathbb{P}$. This means that any typical sequence contains exactly $C(\ell)$ different words of length $\ell$. Let us assume that the process $\mathbb{P}$ is such that all possible $s$-sets, which can be extracted from a typical sequence $x = x_1, x_2, \cdots$, have the same probability of occurring. Denote by $N_{\ell,s}$ the number of different $\ell$-words in the random sample $S$. We want to compute the distribution $f(c) = \mathbb{P}(N_{\ell,s} = c)$, which we assume to be independent of the distance $\Delta$. In order to compute $f(c)$ we proceed as follows. Consider first the total number $\mathcal{N}$ of different $s$-sets containing $c$ different elements (with all necessary repetitions) taken from a collection of cardinality $C(\ell)$. A straightforward computation leads to

$$\mathcal{N} = \binom{s - 1}{c - 1} \binom{C(\ell)}{c},$$

which we obtain by first considering the number of possible subsets of cardinality $c$ taken from a set of total cardinality $C(\ell)$, then computing the number compositions of $s$ objects into exactly $c$. Denote by $\mathcal{M}$ the total number of different subsets of cardinality $s$ that can be made up from a set of $C(\ell)$ different objects (with the possibility of repeating each item in the set as many times as possible). It is easy to see that

$$\mathcal{M} = \binom{C(\ell) + s - 1}{s}.$$

Therefore we have that,

$$f(c) = \frac{1}{\mathcal{M}} \binom{s - 1}{c - 1} \binom{C(\ell)}{c} = \binom{C(\ell) - 1}{s} \left( \binom{C + s - 1}{s} \right)^{-1}.$$  \hspace{1cm} (A.1)

The expected value of $N_{\ell,s}$ is easily obtained as follows,

$$\mathbb{E}(N_{\ell,s}) = \frac{1}{\mathcal{M}} \sum_{c=1}^{s} \binom{s - 1}{c - 1} \binom{C(\ell)}{c}$$

$$= \frac{C}{\mathcal{M}} \sum_{c=1}^{s} \binom{s - 1}{c - 1} \binom{C - 1}{c - 1}$$

$$= C \left( \binom{C + s - 2}{s - 1} \left( \binom{C + s - 1}{s} \right)^{-1} \right) = \frac{Cs}{C + s - 1},$$ \hspace{1cm} (A.2)

where the last sum was obtained from the Chu–Vandermonde identity [47], which establishes that

$$\sum_{x=0}^{r} \binom{k}{r-x} \binom{l}{x} = \binom{k+l}{r}$$ \hspace{1cm} (A.3)

for all $r, k, l \in \mathbb{N}_0$. 

J. Phys. A: Math. Theor. 49 (2016) 445601 R Salgado-García and E Ugalde
To compute the variance of $N_{\ell,s}$, we first calculate the expected value of $N_{\ell,s}(N_{\ell,s} - 1)$,

$$\mathbb{E}(N_{\ell,s}(N_{\ell,s} - 1)) = \frac{1}{\mathcal{M}} \sum_{c=1}^{s} c(c-1) \binom{s-1}{c-1} \binom{C(\ell)}{c}$$

$$= \frac{C(\ell) \binom{C(\ell) - 1}{s-1}}{\mathcal{M}} \sum_{c=1}^{s} \binom{C(\ell) - 2}{c-2}$$

$$= C(\ell) \binom{C(\ell) - 1}{s-2} \binom{C + s - 3}{s}^{-1}. \quad (A.4)$$

Now, since the variance of $N_{\ell,s}$ is given by,

$$\text{Var}(N_{\ell,s}) = \mathbb{E}(N_{\ell,s}(N_{\ell,s} - 1)) + \mathbb{E}(N_{\ell,s}) - (\mathbb{E}(N_{\ell,s}))^2,$$

therefore we have that,

$$\text{Var}(N_{\ell,s}) = \frac{C(\ell) C(\ell) - 1}{(C(\ell) + s - 1)(C(\ell) + s - 2)} \binom{C(\ell)}{s} + \frac{C(\ell) s}{C(\ell) + s - 1} \left( \frac{C^s}{C + s - 1} \right)^2$$

$$= \frac{C(\ell) s}{C(\ell) + s - 1} \left( \frac{(C(\ell) - 1)(s - 1)}{C(\ell) + s - 2} + 1 \right) \left( \frac{C(\ell) s}{C(\ell) + s - 1} \right)^2$$

$$= \frac{C(\ell) s}{C(\ell) + s - 1} \left( \frac{(C(\ell) - 1)(s - 1)}{(C(\ell) + s - 2)(C(\ell) + s - 1)} \right),$$

which is the identity given in equation (3).

The proposed estimator is a function of the random variable $N_{\ell,s}$,

$$c_{\ell,s}(N_{\ell,s}) = \frac{sN_{\ell,s}}{s + 1 - N_{\ell,s}}.$$

The expected value of this estimator is given by

$$\mathbb{E}(c_{\ell,s}) = \frac{1}{\mathcal{M}} \sum_{c=0}^{s} \binom{s}{c-1} \binom{C(\ell)}{c}$$

$$= \frac{C(\ell)}{\mathcal{M}} \sum_{c=0}^{\ell} \binom{s}{c-1} \binom{C(\ell) - 1}{c - 1}$$

$$= \frac{C(\ell)}{\mathcal{M}} \sum_{c=0}^{\ell-1} \binom{s}{c} \binom{C(\ell) - 1}{c}.$$

In the last equality we used the symmetry property for the binomial coefficient. To apply the Chu–Vandermonde identity we add and subtract the term $c = s$ and we obtain

$$\mathbb{E}(c_{\ell,s}) = \frac{C(\ell)}{\mathcal{M}} \sum_{c=0}^{\ell} \binom{s}{c} \binom{C(\ell) - 1}{c} - \binom{C(\ell) - 1}{s} \mathcal{M}^{-1},$$

or equivalently

$$\mathbb{E}(c_{\ell,s}) = C(\ell) - (C(\ell) - s) \binom{C(\ell)}{s} \mathcal{M}^{-1}, \quad (A.5)$$

which proves equality (5).
References

[1] Trifonov E and Sussman J 1980 Proc. Nat. Acad. Sci. USA 77 3816–20
[2] Arneodo A, d’Aubenton Carafa Y, Audit B, Bacry E, Muzy J and Thermes C 1998 Eur. Phys. J. B-Condense. Matter 1 259–63
[3] Buldyrev S, Goldberger A, Havlin S, Mantegna R, Matsa M, Peng C K, Simons M and Stanley H 1995 Phys. Rev. E 51 5084
[4] Peng C K, Buldyrev S, Goldberger A, Havlin S, Sciortino F, Simons M and Stanley H 1992 Nature 356 168–70
[5] Peng C K, Buldyrev S, Havlin S, Simons M, Stanley H E and Goldberger A 1994 Phys. Rev. E 49 1685
[6] Peng C K, Buldyrev S, Goldberger A, Havlin S, Mantegna R, Simons M and Stanley H 1995 Physica 221 180–92
[7] Stanley H, Buldyrev S, Goldberger A, Havlin S, Peng C K and Simons M 1999 Physica 273 1–8
[8] Hsieh L C, Luo L, Ji F and Lee H C 2003 Phys. Rev. Lett. 90 018101
[9] Koroteev M V and Miller J 2011 Phys. Rev. E 84 061919
[10] Li W 1991 Phys. Rev. A 43 5240
[11] Messer P W, Arndt P F and Lässig M 2005 Phys. Rev. Lett. 94 138103
[12] Saakian D 2008 Phys. Rev. E 78 061920
[13] Zaks M 2001 Phys. Rev. E 65 011111–1
[14] Gorman J and Greene E 2008 Nature Struct. & Mol. Biol 15 768–74
[15] Mirny L, Slutsky M, Wunderlich Z, Tafvizi A, Leith J and Kosmrlj A 2009 J. Phys. A: Math. Theor. 42 434013
[16] Slutsky M, Kardar M and Mirny L 2004 Phys. Rev. E 69 061903
[17] Tafvizi A, Mirny L A and van Oijen A M 2011 Chemphyschem 12 1481–9
[18] Ashkenasy N, Sánchez-Quesada J, Bayley H and Ghadiri M R 2005 Angewandte Chemie 117 1425–8
[19] Arneodo A, Vaillant C, Audit B, Argoul F, d’Aubenton Carafa Y and Thermes C 2011 Phys. Rep. 498 45–188
[20] Branton D et al 2008 Nat. Biotechnol. 26 1146–53
[21] Huynen M 1996 J. Mol. Evol. 43 165–9
[22] Gu Z, Steinmetz L, Gu X, Scharf C, Davis R and Li W 2003 Nature 421 63–6
[23] Visser J et al 2003 Evolution 57 1959–72
[24] Ciliberti S, Martin O and Wagner A 2007 Proc. Nat. Acad. Sci. USA 104 13591–6
[25] Whitacre J and Bender A 2010 Degeneracy: a design principle for achieving robustness and evolvability J. Theor. Biol. 263 143–53
[26] Masel J and Trotter M 2010 Trends Genet. 26 406
[27] Wagner A 2000 Nat. Genet. 24 355–61
[28] Wagner A 2008 Proc. R. Soc. B 275 91–100
[29] Lind D A and Marcus B 1995 An Introduction to Symbolic Dynamics and Coding (Cambridge: Cambridge University Press)
[30] Koslicki D 2011 Bioinformatics 27 1061–7
[31] Colosimo A and De Luca A 2000 J. Theor. Biol. 204 29–46
[32] Denker M, Grillenberger C and Sigmund K 1976 Ergodic Theory on Compact Spaces (Berlin: Springer-Verlag) (doi:10.1007/BFb0082364)
[33] Jaeger S, Lima R and Mossé B 2003 Bull. Braz. Math. Soc. 34 457–77
[34] Anisiu M C and Cassaigne J 2004 Revue d’Analyse Numerique et de Theorie de l’Approximation 33 123–40 (http://ictp.acad.ro/anisu/papers/2004-Anisiu-Cassaigne-Properties.pdf)
[35] Vogel H 2013 arXiv:1309.3441
[36] Nei M and Kumar S 2000 Molecular Evolution and Phylogenetics (Oxford, USA: Oxford University Press)
[37] Benson D A, Boguski M S, Lipman D J and Ostell J 1997 Nucleic Acids Research 25 1–6
[38] Hochman M 2011 Ergod. Theor. Dynam. Sys. 32 119–40
[39] Allouche J P 1994 Bull. Belg. Math. Soc. 1 133–43 (http://www.emis.de/journals/BBMS/Bulletin/bull942/ALL-PDF)
[40] Ferenczi S 1999 Discr. Math. 206 145–54
[41] Queffelec M 2010 Substitution Dynamical Systems (Berlin: Springer-Verlag) (doi:10.1007/978-3-642-11212-6)
[42] Li W and Kaneko K 1992 *EPL* (*Europhys. Lett.*) 17 655
[43] Salgado-García R and Ugalde E 2013 *J. Stat. Phys.* 153 842–63
[44] Zhang J 2003 *Trends in Ecology & Evolution* 18 292–8
[45] Taylor J S and Raes J 2004 *Annu. Rev. Genet.* 38 615–43
[46] Massip F and Arndt P 2013 *Phys. Rev. Lett.* 110 148101
[47] Roy R 1987 Binomial identities and hypergeometric series *Am. Math. Mon.* 94 36–46