Dynamics of transposable elements generates structure and symmetries in genetic sequences

Giampaolo Cristadoro, Mirko Degli Esposti, and Eduardo G. Altmann

1Dipartimento di Matematica e Applicazioni, Università di Milano - Bicocca, Via Roberto Cozzi 55, I-20125 Milan, Italy
2Department of Computer Science and Engineering, University of Bologna, Mura Anteo Zamboni 7, 40126 Bologna, Italy
3School of Mathematics and Statistics F07, University of Sydney NSW 2006, Australia

(Received 26 August 2019; accepted 18 August 2020; published 23 September 2020)

Genetic sequences are known to possess nontrivial composition together with symmetries in the frequencies of their components. Recently, it has been shown that symmetry and structure are hierarchically intertwined in DNA, suggesting a common origin for both features. However, the mechanism leading to this relationship is unknown. Here we investigate a biologically motivated dynamics for the evolution of genetic sequences. We show that a metastable (long-lived) regime emerges in which sequences have symmetry and structure interlaced in a way that matches that of extant genomes.

**Introduction.** Transposable elements (TEs) are DNA sequences that can relocate themselves in new sites of the genome. They were first discovered in maize by McClintock in the mid-1940s and initially considered as parasites with no functional roles [1]. Nowadays TEs are known to be ubiquitous in both prokaryotes and eukaryotes genomes [2,3] and little doubts are left of their prominent role in genome evolution, shaping structure and function in a multitude of ways [4,5]. As TEs constitute more than half of the sequence ubiquity of complex structures in genomes [30–38], this result raises the question whether symmetry can appear without a full randomization of the sequence and in a way that is compatible with the existence of structure. The importance of this question is enhanced by our recent findings [39] that Chargaff symmetry extends beyond the frequencies of short oligonucleotides—remaining valid on scales where nontrivial structure is present—and that a hierarchy of other symmetries exists, nested at different structural scales. These findings are confirmed in Fig. 1, which shows how commonly used indicators of structures, such as recurrence-time distribution [Fig. 1(a)] and correlation functions [Fig. 1(b)], coincide for symmetrically related observables at different scales.

In this paper we present a biologically motivated dynamical process that explains the observed relation between symmetry and structure in DNA sequences. In particular, we propose a model that mimics the action (inversions/transpositions) of TEs on DNA and we analytically describe its dynamical behavior. Using indicators to quantify both symmetry and the presence of nontrivial structure in symbolic sequences, we show that the co-occurrence of symmetry and structure is an emergent statistical property in sequences generated by such a model, reproducing the same hierarchical relation detected in extant genomes.

Quantifying structure and symmetry. We consider symbolic sequences $s = \{s_i\}_{i=1}^N$ of length $|s| = N$ with $s_i \in \mathcal{A} = \{A, C, G, T\}$. Given a subsequence $\mathbf{a}$ of $s$ (a word) we denote its corresponding reverse-complemented word as $\hat{\mathbf{a}}$, obtained from $\mathbf{a}$ by reversing the order of the symbols and substituting each nucleotide with its conjugated $A \leftrightarrow T$ and $C \leftrightarrow G$ (e.g., $\mathbf{w} = ACTGGCT$, $\hat{\mathbf{w}} = AGCCAGT$). It has been first observed by Chargaff in the 1950s [13] and since then detected across different organisms leading to different proposals for its origin and function [14–29]. The importance of the Albrecht-Buehler explanation is that it shows how this symmetry naturally emerges as an asymptotic outcome of the cumulative action of inversions/transpositions, one of the main mechanisms of relocation of TEs. As we will show, while the proposed mechanism nicely induces Chargaff symmetry in the asymptotic DNA, it does it at the cost of trivialization of the structural properties of the sequence: Symmetry is obtained because of the complete randomization of the full double-stranded DNA. In view of the ubiquity of complex structures in genomes [30–38], this result raises the question whether symmetry can appear without a full randomization of the sequence and in a way that is compatible with the existence of structure. The importance of this question is enhanced by our recent findings [39] that Chargaff symmetry extends beyond the frequencies of short oligonucleotides—remaining valid on scales where nontrivial structure is present—and that a hierarchy of other symmetries exists, nested at different structural scales. These findings are confirmed in Fig. 1, which shows how commonly used indicators of structures, such as recurrence-time distribution [Fig. 1(a)] and correlation functions [Fig. 1(b)], coincide for symmetrically related observables at different scales.

In this paper we present a biologically motivated dynamical process that explains the observed relation between symmetry and structure in DNA sequences. In particular, we propose a model that mimics the action (inversions/transpositions) of TEs on DNA and we analytically describe its dynamical behavior. Using indicators to quantify both symmetry and the presence of nontrivial structure in symbolic sequences, we show that the co-occurrence of symmetry and structure is an emergent statistical property in sequences generated by such a model, reproducing the same hierarchical relation detected in extant genomes.

**Quantifying structure and symmetry.** We consider symbolic sequences $s = \{s_i\}_{i=1}^N$ of length $|s| = N$ with $s_i \in \mathcal{A} = \{A, C, G, T\}$. Given a subsequence $\mathbf{a}$ of $s$ (a word) we denote its corresponding reverse-complemented word as $\hat{\mathbf{a}}$, obtained from $\mathbf{a}$ by reversing the order of the symbols and substituting each nucleotide $A \leftrightarrow T$ and $C \leftrightarrow G$. We call $f_s(\mathbf{a})$ the percentage of the nucleotide $x$ in the sequence $s$. Finally, we denote by $CG(s) := f_G(s) + f_C(s)$ (the so called CG content). In the following, it will be useful to partition the full set $\mathcal{A}^N$ into disjoint subsets of fixed CG content $B^N(k) := \{s \in \mathcal{A}^N | CG(s) = k/N\}$; $\mathcal{A}^N = \sqcup_{k=0}^{N} B^N(k)$.
organization. We thus quantify structure as the distance of \( s \) from random sequences by

\[
I_{\text{st}}(s) = \frac{1}{4} \sum_{x \in A} \left( \frac{1}{\sqrt{1-f_x(s)}} \sigma_x(x) - 1 \right),
\]

where \( \mu_x \equiv \langle \tau \rangle \) and \( \sigma_x \equiv \sqrt{\langle \tau^2 \rangle - \langle \tau \rangle^2} \) are the mean and standard deviation of the measured \( P(\tau) \), and \( \sqrt{1-f_x} \) is the expected \( \sigma_x/\mu_x \) for nucleotide \( x \) in a random sequence. For random sequence we thus have \( I_{\text{st}}(s) = 0 \), while departure from this value mark the presence of nontrivial structure. For simplicity, we consider \( I_{\text{st}} > 0.01 \) to be a signature of structure.

**Dynamics.** We investigate symmetry and structure of sequences that evolve through the following dynamics, that maps one sequence \( s(t) \in A^N \) into another sequence \( s(t+1) \in A^N \) by mimicking the action of TEs [12]. The dynamics is defined composing two actions:

(i) Pick a random position \( j \) of \( s \) and a random size \( \ell \geq 0 \), with \( (\ell) = L [51] \).

(ii) Replace the subsequence \( b \equiv \{s_i\}_{i=j}^{j+\ell-1} \) of size \( \ell \) starting at position \( j \), by its reverse complement \( b \).

The couple \( (j, \ell) \) parametrizes the effect of an inversions/translations, which we denote by \( B_{(j,\ell)} : A^N \rightarrow A^N \). Its action has interesting properties: \( B_{(j,\ell)} \) is an involution for every \( j, \ell \) and the total number of \( C \) and \( G \) (or, equivalently, of \( A \) and \( T \)) is invariant under \( g: CG(s_0) = CG(s_{0 \ell}) \) \( \forall t \). This implies that the dynamics is restricted to the invariant subspace of sequences with constant CG content \( B^N[CG(s_0)] \).

**Asymptotic equilibrium.** The dynamics can be equivalently described as an ergodic Markov chain over the space of sequences \( B^N[CG(s_0)] \). The fact that \( B_{(j,\ell)} \) is an involution forces the transition matrix to be bi-stochastic and thus in the asymptotic equilibrium all sequences are equiprobable. This means that, for \( t \rightarrow \infty \) and irrespective of the initial ancient DNA sequence, the evolution asymptotically leads to sequences that can be equivalently considered generated by an independent and identically distributed (iid) process with \( p(G) = p(C) = CG(s_0)/2 \) and \( p(A) = p(T) = (1 - CG(s_0))/2 \). Therefore, the expected value of our indicators of symmetry and structure Eqs. (1) and (2) vanish asymptotically,

\[
\lim_{t \rightarrow \infty} I_{st}(s(t)) = \lim_{t \rightarrow \infty} I_{\text{sym}}(s(t)) = 0,
\]

for any initial sequence \( s(0) \) [52]. This shows analytically that the TE dynamics asymptotically leads to Chargaff symmetric sequences, in agreement with previous claims [12]. However, this symmetric equilibrium is a (trivial) consequence of a full randomization. Therefore our results show also that the current explanation of the second Chargaff parity rule [12] is not satisfactory as it is not compatible with any structure, which is known to remain significant at distances of several thousands of nucleotides [30–38] (see also Fig. 1). Next we show that the same TE dynamics is rich enough by showing that symmetric sequences with nontrivial structure are generated pre-asymptotically as long-lived metastable states of TEs dynamics.

**Symmetry and structure over time—three regimes.** We now investigate symmetry and structure of the sequences \( s(t) \) by

![FIG. 1. Symmetry and structure are intertwined in DNA. Results are shown for Homo sapiens chromosome 1 (symbols) and its randomly shuffled version (dashed lines). Each curve corresponds to one observable. Symmetrically related observables appear in the same box in the legend. (a) Distribution \( P(\tau) \) of recurrence times \( \tau \) (measured in number of bases) between successive occurrences of the same nucleotide and its symmetric one (see [21] where a similar measure was first introduced). (b) Probability \( f_x(X_A, X_B) \) that the bigrams \( X_A \) and \( X_B \) appear separated by a distance \( \ell \), Plotted is the normalized cross-correlation \( \tilde{z}(X_A, X_B) = f_{X_A, X_B}/(f_A, f_B) \) as a function of \( \ell \), for symmetrically related couples \( [X_A, X_B] \) (see legend). Different nested symmetries are valid at different scales \( \ell \) (see Ref. [39] for further details): for \( \ell \leq 150 \) Chargaff \( z(X_A, X_B) = z(X_B, X_A) \), for \( 150 \leq \ell \leq 1500 \) Chargaff and reverse symmetry \( z(X_A, X_B) = z(X_B, X_A) \), and for \( \ell \geq 1500 \) complement \( z(X_A, X_B) = z(X_A, X_B) \) and reverse symmetry.

\( \langle x \rangle \)}
computing how our indicators $I_{\text{sym}}$ an $I_{\text{st}}$ depend on time $t$ (i.e., their values after $t$ applications of $g_{t}(t)$). We show that Chargaff symmetry emerges much before equilibrium, together with a complex domainlike structure.

We first investigate structural properties of sequences after a finite number $t$ of iterations. We define a domain of $s(t)$ as a subsequence of consecutive sites that have been involved in the same series of reverse-complement events. We then distinguish between domains of type $\Gamma$ and $\bar{\Gamma}$, depending on whether the number of transformations $g$ they were involved is even or odd, respectively. By definition, the starting sequence is composed by a single domain of type $\Gamma$. After one iteration it is split into three domains, two of type $\Gamma$ and one of type $\bar{\Gamma}$ of length $\ell_1$, corresponding to the subsequence involved in the first reverse-complement event. We now compute the average sizes $\langle \ell_1 \rangle (t)$ and $\langle \ell_2 \rangle (t)$ of domains after $t$ iterations. Three regimes can be identified.

(i) For short times $t$, if $L \ll N$, the probability that the first few iterates all involve different subsequence is very high [53]. At each iterate, a subsequence of a domain of type $\Gamma$ occurs with probability $p = 0.5$. We have that in this regime:

$$\langle \ell_1 \rangle (t) = L \quad \text{and} \quad \langle \ell_2 \rangle (t) = N/t.$$  

This regime lasts until iterates start overlapping, which happens when $N/t \approx L$ and average domain sizes equalize

$$\langle \ell_1 \rangle (t) = \langle \ell_2 \rangle (t) = L.$$  

This regime is thus valid for $0 < t \lesssim t_{\text{metastable}} = N/L$.

(ii) For $t \gtrsim t_{\text{metastable}} = N/L$ a typical reverse-complement event will overlap with one domain. In this case all the domains that lie fully inside the subsequence involved in the reverse-complement event will change type (and position) without changing length; the domains at the border are instead split in two subdomains of different types. The randomness of this process guarantees that the already reached balance between the number and average length of the two domain types $\Gamma$ and $\bar{\Gamma}$ is not broken while their common average length decreases in time as

$$\langle \ell_1 \rangle (t) = \langle \ell_2 \rangle (t) = N/t.$$  

This second regime ends after a number of iterations $t \sim t_{\text{equilibrium}} = N$ when equilibrium is reached.

(iii) For $t > t_{\text{equilibrium}} = N$ the average lengths stabilize at the stationary value,

$$\langle \ell_1 \rangle (t) = \langle \ell_2 \rangle (t) = 1,$$  

and the sequence can be thought as a realization of the asymptotic equilibrium discussed above.

We now explain how structure $I_{\text{st}}(s(t))$ and symmetry $I_{\text{sym}}(s(t))$ depend on the domain sizes $\langle \ell_1 \rangle$ and $\langle \ell_2 \rangle$ and thus on the different regimes.

$I_{\text{st}}(s)$: In order to identify the contribution of the dynamics in generating complex structural features, we consider an initial $s(0)$ generated by an iid process [no structure, $I_{\text{st}}(0) = 0$]. With this choice, a value $I_{\text{st}} \neq 0$ signals the construction, under the action of the dynamics, of different domain types. In particular, at $t_{\text{metastable}}$ and for $L \gg 1$, the total variance $\sigma^2$ can be estimated, using the law of total variance, as the sum of two components: one that measures variability of the mean of returns between domain types and the other measuring variability of returns within each type. Accordingly $I_{\text{st}}(t)$ grows from 0 to the value $I_{\text{st}}(t_{\text{metastable}}) > 0$ at the end of the first regime. In the second regime the domain sizes decay and $I_{\text{st}}(t)$ decreases to zero at equilibrium (at $t_{\text{equilibrium}}$). In terms of regimes we thus expect (i) $I_{\text{st}}$ grows; (ii) $I_{\text{st}}$ decays; (iii) $I_{\text{st}} = 0$.

$I_{\text{sym}}(s)$: Each domain of type $\Gamma$ is a subsequence of the ancient sequence $s(0)$. If average size of such domains at time $t$ is large enough, the frequencies of each nucleotide are approximately the same as their frequency in $s(0)$; similarly for $\bar{\Gamma}$ and sym $(s)$. No constraints are imposed to the symmetry of the ancient genome. In particular, if the original sequence is not Chargaff symmetric $I_{\text{sym}}(s(0)) > 0$ then the symmetry remains broken for all $t \lesssim t_{\text{metastable}}$, as quantified by $I_{\text{sym}}(s(t)) \geq \frac{1}{2} (\langle \ell_1 \rangle (t) - \langle \ell_2 \rangle (t)) I_{\text{sym}}(s(0))$. In terms of regimes we thus expect (i) $I_{\text{sym}} > 0$; (ii) $I_{\text{sym}} = 0$; (iii) $I_{\text{sym}} = 0$.

Altogether, the estimations and calculations above lead to the following predictions for the presence of symmetry and structure as a function of time $t$ (regimes i–iii).

(i) $0 \leq t \leq t_{\text{metastable}} = N/L$:

- Structure $I_{\text{st}} > 0$ but no symmetry $I_{\text{sym}} > 0$.

(ii) $t_{\text{metastable}} = N/L \leq t \leq t_{\text{equilibrium}} = N$:

- Structure $I_{\text{st}} > 0$ and symmetry $I_{\text{sym}} = 0$.

(iii) $t_{\text{equilibrium}} = N < t$:

- Symmetry $I_{\text{sym}} = 0$ but no structure $I_{\text{st}} = 0$.

In Fig. 2 we confirm these predictions in a numerical simulation.
The metastable regime. The crucial feature of the TE dynamics discussed above is that in regime (ii) both nontrivial structure and symmetry coexist in the generated sequences. The time (measured in number of iterations) for which this regime is valid is orders of magnitude larger than that of the first regime, as the ratio $t_{\text{equilibrium}}/t_{\text{metastable}} \simeq L$ corresponds to the average size of transposable elements (for example, $L \simeq 10^5$ in Homo sapiens [46]). We thus denote such long-lived regime as metastable and we expect it to be generically observed, even though it does not correspond to the stable equilibrium of our model.

The DNA sequences in the metastable regime are characterized by a symmetric domainlike structure. Domain models have been already introduced in literature to reproduce the complex structure generically observed in extant DNAs [11,40–46]. In particular if the distribution of domain sizes has a fat tail, this will lead to a long-range correlated sequence [11], signaled by a slow decay of $P(\tau)$. The novelty of our approach is twofold: First, the domainlike structure in the metastable regime is an emergent property of the TE dynamics (it is not imposed a priori); secondly, such complex structure is intertwined with symmetry, that itself is an output of the dynamics. In particular, we have shown that sequences in the metastable regime are not only Chargaff symmetric ($I_{\text{sym}} = 0$), they reproduce the hierarchical relation between symmetry and structure that is a distinctive feature of extant genomes (see Fig. 3).

Different organisms. In Fig. 4 we report $I_{\text{sym}}$ and $I_{\text{str}}$ computed for genomes of different families, together with the values obtained from our dynamics. It shows that symmetry and structure coexist in most cases. The sequences from animals show enhanced structure while the cases of archaea and bacteria show a moderate signature of structure, in agreement with the temporal behavior of our model (i.e., associating $t$ with the age of the genomes). Note that symmetry and structure properties are both statistical observations we made on the full DNA sequence. Any evolutionary constraint that pertains a small percentage of an organism genome does not affect these statistical observations in a sensible way. As an example, the protein-coding regions of Homo sapiens account for 1.5% of the full sequence. On the other hand, care should be taken when dealing with many different organisms: Extensions of the model incorporating additional aspects of DNA evolution will be required for a quantitative comparison with the empirical data.

Conclusion. We have shown how a model that captures the action of transposable elements (TEs) is able to reproduce the intricate relation between symmetry and structure present in DNA sequences. We find that symmetry and structure change differently at different time scales (i.e., for different number of actions of TEs). For a large (pre-asymptotic) time interval, the sequences obtained in our model show the same nontrivial structures and a hierarchy of symmetries (including Chargaff) as in actual DNA sequences [see Figs. 1(b) and 3(b)]. Our mathematical model is extremely simplified and includes the essential elements to explain the onset of symmetry and structure. In particular, it mimics only a simple action of TEs (reverse-complement), ignoring the fact that TEs are classified in different families, have different properties, and act according to different mechanisms [47–49]. We expect that incorporating more details of the TE dynamics in our model will refine our understanding of their role in shaping statistical properties of DNA sequences, in particular in an evolutionary viewpoint that would lead to refinements in the data-model comparison presented in Fig. 4.
DYNAMICS OF TRANSPOSABLE ELEMENTS GENERATES ... PHYSICAL REVIEW RESEARCH 2, 033477 (2020)


[51] More precisely, the pairs \((j, \ell)\) are drawn, independently from previous iterations, from a joint distribution \(\rho(j, \ell)\) chosen such that its marginal \(\phi(\ell)\) has support contained in \([0, N]\), finite average \(L\), and the conditional distribution of positions \(\psi(j|\ell)\) is uniform in \([1, N - \ell + 1]\). We consider distributions \(\phi(\ell)\) that guarantee ergodicity of the Markov Chain. We expect ergodicity to be generically valid; e.g., it suffices to have non-zero probability for the identity transformation (i.e. \(\phi(0) \neq 0\) and for single-nucleotide complementing (\(\psi(1) \neq 0\).

[52] We can intuitively understand this result by noting that each action of the transposon effectively creates two cuts in the sequence and moves them by a distance \(L\) on average. Since cuts can happen at any location, this process eventually mixes complementary basis at different positions and breaks any correlations originally present in \(s(0)\).

[53] Quantitatively, if we drop \(t \leq N/L\) points uniformly at random on an interval of length \(N\), they will be separated by a distance at least \(L\) with probability \(1 - (t - 1)L/N\).