ISSN 1560-3547, Regular and Chaotic Dynamics, 2018, Vol. 23, No. 2, pp. 135–151. © Pleiades Publishing, Ltd., 2018.

Distribution of Base Pair Alternations in a Periodic DNA Chain: Application of Pólya Counting to a Physical System

Malcolm Hillebrand[1*], Guy Paterson-Jones[1], George Kalosakas[2***], and Charalampos Skokos[1****]**

[1] *Department of Mathematics and Applied Mathematics, University of Cape Town, Rondebosch, Cape Town 7701, South Africa*

[2] *Department of Materials Science, University of Patras, Rio GR-26504, Greece*

Received October 13, 2017; accepted December 11, 2017

Abstract—In modeling DNA chains, the number of alternations between Adenine–Thymine (AT) and Guanine–Cytosine (GC) base pairs can be considered as a measure of the heterogeneity of the chain, which in turn could affect its dynamics. A probability distribution function of the number of these alternations is derived for circular or periodic DNA. Since there are several symmetries to account for in the periodic chain, necklace counting methods are used. In particular, Pólya's Enumeration Theorem is extended for the case of a group action that preserves partitioned necklaces. This, along with the treatment of generating functions as formal power series, allows for the direct calculation of the number of possible necklaces with a given number of AT base pairs, GC base pairs and alternations. The theoretically obtained probability distribution functions of the number of alternations are accurately reproduced by Monte Carlo simulations and fitted by Gaussians. The effect of the number of base pairs on the characteristics of these distributions is also discussed, as well as the effect of the ratios of the numbers of AT and GC base pairs.

MSC2010 numbers: 05A15, 92D20

DOI: 10.1134/S1560354718020016

Keywords: DNA models, Pólya's Counting Theorem, heterogeneity, necklace combinatorics

1. INTRODUCTION

Single circular DNA molecules are abundant in nature. The whole genome in a typical bacterium is usually contained in a closed DNA molecule, while in eucaryotes the organelle DNA, inside the mitochondria and chloroplasts, is also found in the same form [1, 23]. Also plasmids, either naturally found in bacteria, or used as vectors in gene cloning, are smaller circular DNA segments. Apart from these cases, in considering the dynamics and other properties of DNA chains, it is often useful to model the chain using periodic boundary conditions in order to avoid finite size or edge effects. For example, periodic boundary conditions have been used to study denaturation bubbles and the melting behavior of DNA [2, 6, 13, 37, 39, 43], probability distributions of thermal openings in the double strand [7, 18], bubble opening profiles in promoter regions which regulate gene transcription [3–5, 11, 12, 16, 20], binding sites of DNA-associated proteins [26, 38], various dynamical and nonlinear properties of DNA [21, 27, 28, 40, 41, 44], as well as charge transport in DNA [10, 14, 17, 19, 33].

A DNA chain consists of a series of base pairs, where each base pair is either Adenine–Thymine (AT) or Guanine–Cytosine (GC). Currently, we are investigating the influence of different factors on the chaoticity of periodic DNA chains [36]. One of the examined quantities is the number of

*E-mail: malcolm.hillebrand@gmail.com
**E-mail: guy.paterson.jones@gmail.com
***E-mail: georgek@upatras.gr
****E-mail: haris.skokos@uct.ac.za

base pair alternations, which can be considered as a quantifier of the system's heterogeneity. In this work we focus on the rigorous mathematical treatment of alternation counting in periodic DNA sequences. To study periodic DNA, we will consider the DNA necklace associated to a DNA chain, where the first and the last base pairs in the chain will become neighbors. This periodicity presents some modeling challenges — if one considers two distinct chains of DNA, it may still be the case that their corresponding necklaces are the same, as one may be merely a rotation or reflection of the other. Such symmetries need to be addressed if any conclusions are to be made about the structure and the dynamics of DNA necklaces. In particular, we are concerned with the number α of base pair alternations in the necklace, where an alternation is defined to be a point at which an AT base pair neighbors a GC base pair or vice versa. Consider, for instance, the DNA chain shown in Fig. 1. Representing a GC base pair (black bead) with a 0 and an AT base pair (white bead) with a 1, the chain can be written in the form $(1)000\bar{1}0\bar{1}1\bar{0}0\bar{1}(0)$. Here, we have given the leftmost base pair at each alternation point an overbar, and used brackets to denote the fact that in the corresponding DNA necklace the first and last base pairs are neighbors. This necklace is illustrated in Fig. 2, and counting the number of overbars we see that there are $\alpha = 6$ alternations.

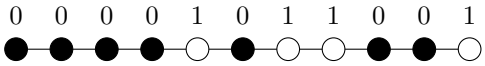

Fig. 1. An example of a DNA chain. GC base pairs are represented by black beads and the number 0, while AT base pairs are represented by white beads and the number 1. In the DNA necklace corresponding to this chain, the AT base pair at the far right neighbors the GC base pair at the far left.

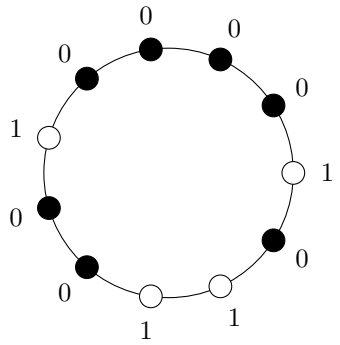

Fig. 2. The DNA necklace corresponding to the chain of Fig. 1. This necklace has $\alpha = 6$ alternations.

It is worth noting that a base pair alternation corresponds to the appearance of the particular sequences (often referred to as "words") 01 or 10 in a DNA chain. Word occurrence probabilities have already been studied in the literature (see, e. g., [22, 24, 30–32, 34, 35] and references therein), with emphasis on the appearance of patterns with unexpectedly high or low frequencies, as well as on repeating sequences. However, these studies concern the case of linear DNA segments, or, in other words, DNA chains with fixed boundary conditions. The periodic boundary conditions we consider in our study make the problem of counting alternations (or, more generally, the appearance of specific words) in circular DNA segments much more complicated than in the case of linear DNA segments due to the appearance of additional symmetries in the DNA structures imposed by rotations and/or reflections.

Each base pair in a DNA necklace can contribute at most 2 alternations, depending on which neighbors it differs from. Supposing that the number of AT and GC base pairs in the necklace is given by N_{AT} and N_{GC}, respectively, this yields the restriction $0 \leqslant \alpha \leqslant \min\{2N_{AT}, 2N_{GC}\}$. We note that in the extreme case of a homogeneous chain composed of base pairs of the same kind $\alpha = 0$, while if both types of base pairs are present in the DNA chain, the smallest possible value of alternations is $\alpha = 2$. The later corresponds to a chain having all AT (and consequently GC) base pairs grouped together. Furthermore, if we traverse the necklace pair by pair until we end up where we started, we must necessarily switch between AT and GC base pairs an even number of times. Thus, $\alpha = 2M$ for some $M \in \mathbb{N}$.

Now the natural question is: what is the probability that a random DNA necklace with a specified number of AT and GC base pairs, N_{AT} and N_{GC}, respectively, has a specified number of alternations α? Or, in other words, how many possible combinations of such base pairs are there that yield α alternations once the cyclic and reflective symmetries are taken into account? In what follows we answer these questions and provide an algorithm for computing the number of distinct DNA necklaces satisfying these constraints.

The paper is organized in the following way: In Section 2, the mathematical background is laid out, leading into a Pólya Enumeration Theorem for bipartite sets. In Section 3 an explicit algorithm for calculating the number of distinct DNA necklaces with given values of α, N_{AT} and N_{GC} is described, while in Section 4 we compare the theoretical results to those obtained from Monte-Carlo simulations and investigate the effect of the N_{AT} and N_{GC} values on the characteristics of the probability distribution function (pdf) of α. Finally, in Section 5 we summarize our results, while in the Appendix we provide a Python computer code implementing the algorithm of Section 3.

2. THEORETICAL TREATMENT

Our problem can be neatly related to the combinatorics of necklaces. Effectively, we are interested in the number of distinct necklaces with $N = N_{AT} + N_{GC}$ beads, where N_{AT} of the beads are white, N_{GC} of the beads are black, and there are α alternations between the colors. We consider necklaces to be the same if they can be reflected or rotated into one another, and beads of the same color are treated as indistinguishable. Because of this, we can equivalently think of a necklace with α alternations as a necklace of α *containers*, where each container carries some number of black or white beads of the same color, and adjacent containers have different colors. This idea is illustrated in Fig. 3.

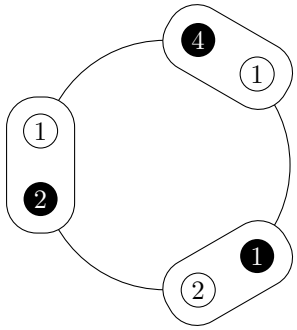

Fig. 3. The necklace of containers corresponding to the DNA necklace of Fig. 2. The numbers in each container represent the number of consecutive black or white beads in that segment of the necklace.

We will refer to containers carrying black beads as black containers, and similarly for white containers. Counting the number of distinct necklaces with the given constraints can thus be reformulated as the problem of assigning numbers of beads to α containers, such that the total of the numbers in the black and white containers is equal to N_{GC} and N_{AT}, respectively. Two such assignments will be considered equivalent if the containers can be rotated or reflected into one another in such a way as to preserve both the colors and numbers of beads they contain.

Enumerating such assignments is simpler than enumerating necklaces, as we have one less constraint — the number of alternations is now implicit in the formulation of the problem. To perform this enumeration we will require some tools from Pólya counting theory — in particular, we will need a version of the Pólya Enumeration Theorem for sets partitioned into two parts, which we will refer to as *bipartite* sets. For completeness' sake, we present this material below.

2.1. Group Actions

Let A be a set. Then we define the *symmetric group on* A to be the set of permutations of A:

$$S_A = \{\varphi : A \to A \mid \varphi \text{ is a bijection}\}. \tag{2.1}$$

A *cycle* is a permutation $\varphi \in S_A$ such that there exist distinct elements $\{x_1, x_2, \ldots, x_k\} \in A$ and:

$$\varphi(x) = \begin{cases} x_{i+1} & \text{if } x = x_i \text{ for some } 1 \leqslant i < k \\ x_1 & \text{if } x = x_k \\ x & \text{otherwise.} \end{cases} \tag{2.2}$$

We denote such a cycle suggestively as $(x_1\ x_2\ \ldots\ x_k)$, and say that $\varphi \in S_A$ is a *k-cycle* if $\varphi = (x_1\ x_2\ \ldots\ x_k)$ for some $x_i \in S_A$. Two cycles $(x_1\ x_2\ \ldots\ x_k)$ and $(y_1\ y_2\ \ldots\ y_l)$ are said to be *disjoint* if the sets $\{x_1, x_2, \ldots, x_k\}$ and $\{y_1, y_2, \ldots, y_l\}$ are disjoint.

If A is a finite set, every element of S_A can be written as a composition of cycles; in general, however, this cannot be done uniquely. On the other hand, we have the following fundamental structure theorem for elements of finite symmetric groups (see, for example, [15]):

Theorem (Cycle Decomposition Theorem). *If A is a finite set, then every element $\varphi \in S_A$ can be written as a product of pairwise disjoint cycles, unique up to order of the cycles:*

$$\varphi = (x_{11}\ x_{12}\ \ldots\ x_{1k_1}) \cdots (x_{n1}\ x_{n2}\ \ldots\ x_{nk_n}).$$

Given a group G and a set A, a *group action* of G on A is a homomorphism $\Gamma_G : G \to S_A$. In other words, elements of G are identified with permutations of A in a manner that preserves the group structure. To simplify the notation, we will write gx instead of $\Gamma_G(g)(x)$ for the action of $g \in G$ on some $x \in A$.

The *orbit* of an element $x \in A$ under the group action Γ_G is defined to be the set $\mathrm{Orb}_x = \{gx \mid g \in G\}$, and its *stabilizer* is given by the subgroup $\mathrm{Stab}_x = \{g \in G \mid gx = x\}$. Given some $g \in G$, we denote its set of fixed points by $\mathrm{Fix}_g = \{x \in A \mid gx = x\}$.

2.2. Pólya's Counting Theory

One can often rephrase counting problems in terms of computing the number of distinct orbits of some group action. Pólya's counting theory can be thought of as a tool for making these computations systematic and expedient. A fundamental lemma on which this theory is built is the following [9]:

Lemma 1 (Burnside's Lemma). *The number of distinct orbits in a group action of a finite group G on A is given by the average number of fixed points of elements of G:*

$$\#Orbits = \frac{1}{|G|} \sum_{g \in G} |Fix_g|. \tag{2.3}$$

A basic problem in combinatorics is the following. Suppose one has a finite set of objects A, and one wishes to color them with colors from another set Ω. How many distinct ways are there of coloring the objects up to some kind of symmetry? This can be recast in the language of group actions. The set of possible colorings is given by $\Omega^A = \{\varphi : A \to \Omega \mid \varphi \text{ a function}\}$, and the symmetry is given by a group action Γ_G on A. This group action passes naturally to a group action $\tilde{\Gamma}_G$ on Ω^A, defined by $g\varphi : x \mapsto \varphi(gx)$.

The question now reduces to counting the number of distinct orbits of this latter action. In this simplified case, Burnside's lemma is often sufficient to answer the question. We can generalize this problem slightly, however. Suppose that each color has an associated *weight* given by a function $\omega : \Omega \to \mathbb{N}$. Given a coloring $\varphi : A \to \Omega$ of the objects, we define its *total weight* to be the sum:

$$|\varphi| = \sum_{x \in A} \omega \circ \varphi(x). \tag{2.4}$$

How many distinct colorings of A with a given total weight are there, up to symmetries given by some group action Γ_G? Note that the total weight of any coloring in a given orbit is the same, as elements of g merely permute the set A. Thus, the problem boils down to calculating the number of distinct orbits with a given total weight. Pólya identified two necessary ingredients for a systematic answer to this question: generating functions, and an understanding of the cycle structure of elements of G [29].

Definition (Generating Function). *Let $\omega : \Omega \to \mathbb{N}$ be an assignment of weights to some set Ω. Suppose further that there are at most a finite number of elements of any given weight, that is, $|\omega^{-1}(n)|$ is finite for every $n \in \mathbb{N}$. Then the* generating function *of ω is given by the polynomial:*

$$f_\omega(x) = \sum_{i=0}^{\infty} |\omega^{-1}(i)|\ x^i. \tag{2.5}$$

Generating functions are useful as they encode combinatorial data — in this case the number of colors of a given weight — as algebraic objects. In particular, we will need the following lemma:

Lemma 2. *Let $\omega_1 : \Omega_1 \to \mathbb{N}$ and $\omega_2 : \Omega_2 \to \mathbb{N}$ be assignments of weights to the sets Ω_1 and Ω_2, respectively. Define an assignment of weights to the set $\Omega_1 \times \Omega_2$ by $\omega : (x_1, x_2) \mapsto \omega_1(x_1) + \omega_2(x_2)$. Then $f_\omega(x) = f_{\omega_1}(x) \cdot f_{\omega_2}(x)$.*

Given a group action Γ_G and an element $g \in G$, we denote by $C_k(g)$ the number of k-cycles in the unique disjoint cycle decomposition of $\Gamma_G(g)$. We can now encode information about the cycle structure of elements of G in the following multivariate polynomial:

Definition (Cycle Index). *Let G be a finite group. Then the* cycle index *of a group action Γ_G on a finite set A of cardinality n is given by the polynomial [8]:*

$$Z_G(x_1, x_2, \ldots, x_n) = \frac{1}{|G|} \sum_{g \in G} x_1^{C_1(g)} x_2^{C_2(g)} \ldots\ x_n^{C_n(g)}. \tag{2.6}$$

This cycle index will allow us to efficiently compute the number of distinct orbits of the group action. With this in mind, we are now in a position to state a version of the Pólya counting theorem, answering the generalized problem given earlier:

Theorem (Pólya Enumeration Theorem). *Let A be a finite set of objects, Ω a set of colors, $\omega : \Omega \to \mathbb{N}$ an assignment of weights to the colors with generating function f_ω, and Γ_G a group action of a finite group G on A. Then Γ_G passes naturally to a group action $\tilde{\Gamma}_G$ on Ω^A, and a generating function by total weight for the number of distinct orbits of $\tilde{\Gamma}_G$ is given by*

$$Orbits_{\tilde{\Gamma}_G}(x) = Z_G\big(f_w(x), f_w(x^2), \ldots, f_w(x^n)\big). \tag{2.7}$$

2.3. Pólya Enumeration Theorem for Bipartite Sets

By considering multivariate generating functions, the Pólya enumeration theorem can be generalized to the case where the colors take weights in $\mathbb{N}^k$. We will generalize the theorem in a different direction, however. Suppose we have a partition of A into two parts, $A = X \sqcup Y$, and a group action Γ_G on A. We would like to consider the problem of counting distinct colorings of A under this symmetry, with the additional constraint that we color elements of X from a set Ω_X, and elements of Y from a set Ω_Y. To this end, we will say that a coloring $\varphi : A \to \Omega_X \sqcup \Omega_Y$ is *valid* if $\varphi(x) \in \Omega_X \iff x \in X$ and $\varphi(x) \in \Omega_Y \iff x \in Y$.

There is an obstruction to this, however — the group action may map elements in X to elements in Y or vice versa. In this case, the extension of Γ_G to the set of possible colorings is no longer well-defined, as there is no natural way to compare the sets of colors Ω_X and Ω_Y. Fortunately, this is the only obstruction to proving a Pólya-type theorem for this problem. This motivates the following definition:

Definition (Partition-Preserving Group Action). *Let $A = X \sqcup Y$, and let Γ_G be a group action on A. Then we say that Γ_G is* partition-preserving *if for every $g \in G$, $gx \in X \iff x \in X$ and $gx \in Y \iff x \in Y$.*

The importance of this property is as follows. Suppose we have a group action Γ_G on $A = X \sqcup Y$, and some element $g \in G$. Then $\Gamma_G(g)$ has a unique disjoint cycle decomposition given by $\Gamma_G(g) = C_1 \cdot C_2 \cdot \ldots \cdot C_k$. If Γ_G is partition-preserving, then each cycle C_i is contained entirely in either X or Y, and Γ_G is in fact partition-preserving if and only if this is the case for every $g \in G$.

If Γ_G is partition-preserving, then we define $C_k^X(g)$ to be the number of k-cycles in the disjoint cycle decomposition of $\Gamma_G(g)$ that are contained in X, and we define $C_k^Y(g)$ analogously. We will now define an analogue of the cycle index polynomial for the case of partition-preserving group actions. This will allow us to keep track of the cycle structure of elements of the group as well as which partition part each cycle acts on:

Definition (Bipartite Cycle Index). *Let G be a finite group and $A = X \sqcup Y$ a finite set of cardinality n. Then the* bipartite cycle index *of a partition-preserving group action Γ_G on A is defined to be the polynomial:*

$$\tilde{Z}_G(x_1, \ldots, x_n, y_1, \ldots, y_n) = \frac{1}{|G|} \sum_{g \in G} {x_1}^{C_1^X(g)} \ldots \ {x_n}^{C_n^X(g)} {y_1}^{C_1^Y(g)} \ldots \ {y_n}^{C_n^Y(g)}. \tag{2.8}$$

We can now generalize Pólya's theorem to the case of partition-preserving group actions. We note that this theorem is used implicitly in [29] without proof.

Theorem 1 (Bipartite Pólya Enumeration Theorem). *Let Γ_G be a partition-preserving group action of a finite group G on a finite set $A = X \sqcup Y$. Let $\Omega = \Omega_X \sqcup \Omega_Y$ be a set of colors, and let $\omega_X : \Omega_X \to \mathbb{N}^+$ and $\omega_Y : \Omega_Y \to \mathbb{N}^+$ be their assigned weights with respective generating functions f_X and f_Y. If Φ is the set of valid colorings of A, then Γ_G passes naturally to a group action $\tilde{\Gamma}_G$ on Φ, and a generating function by total weight for the number of orbits of $\tilde{\Gamma}_G$ is given by*

$$Orbits_{\tilde{\Gamma}_G}(x) = \tilde{Z}_G \left(f_X(x), \ldots, f_X(x^k), f_Y(x), \ldots, f_Y(x^k) \right). \tag{2.9}$$

Proof. We pass to a group action $\tilde{\Gamma}_G$ on Φ as follows. Given a valid coloring $\varphi \in \Phi$ and an element $g \in G$, we define the action of g on φ by $g\varphi : x \mapsto \varphi(gx)$. To compute a generating function for the number of orbits of $\tilde{\Gamma}_G$ by total weight, we will determine the generating functions for the number of fixed points of each $g \in G$ by total weight.

Consider some $g \in G$. As A is finite, there exists a unique disjoint cycle decomposition $\Gamma_G(g) = C_1 \cdot C_2 \cdot \ldots \cdot C_k$, where each C_i is a cycle in the symmetric group S_A. Now suppose that g fixes some valid coloring $\varphi \in \Phi$; that is, $g\varphi = \varphi$. Then, assuming the cycle $C_i = (x_1 \ x_2 \ \ldots \ x_{k_i})$ for some $x_i \in A$, we have by definition that $\varphi(x_i) = (g\varphi)(x_i) = \varphi(gx_i) = \varphi(x_{i+1})$, and hence every element in the cycle must have the same color under φ. The number of colorings of C_i that are fixed by g is thus given by the generating function $f_X(x^{k_i})$ if C_i lies in X, and $f_Y(x^{k_i})$ if C_i lies in Y. We note that one of these two cases must occur for every cycle as Γ_G is partition-preserving. By Lemma 2, then, the number of valid colorings of A that are fixed by g is given by the generating function:

$$\mathrm{Fix}_g(x) = f_X^{C_1^X(g)}(x) \cdots f_X^{C_k^X}(x^k) f_Y^{C_1^Y}(x) \cdots f_Y^{C_k^Y}(x^k). \tag{2.10}$$

By Burnside's lemma, the number of orbits of $\tilde{\Gamma}_G$ of a particular weight is given by the average number of fixed colorings of that weight by elements $g \in G$. Applying Burnside's lemma for each possible weight, the number of orbits of $\tilde{\Gamma}_G$ is thus given by the generating function:

$$\begin{aligned} \mathrm{Orbits}_{\tilde{\Gamma}_G}(x) &= \frac{1}{|G|} \sum_{g \in G} \mathrm{Fix}_g(x) \\ &= \frac{1}{|G|} \sum_{g \in G} f_X^{C_1^X(g)}(x) \ldots f_X^{C_k^X}(x^k) f_Y^{C_1^Y}(x) \ldots f_Y^{C_k^Y}(x^k) \\ &= \tilde{Z}_G \left(f_X(x), \ldots, f_X(x^k), f_Y(x), \ldots, f_Y(x^k) \right). \end{aligned} \tag{2.11}$$

□

We note that, as a corollary of this proof, we can recover a bivariate generating function from this expression, where the coefficient of $x^a y^b$ represents the number of distinct colorings with total weight a in Ω_X, and total weight b in Ω_Y:

Corollary. *A bivariate generating function by total weight in Ω_X and Ω_Y, for the number of distinct colorings of A, is given by*

$$Orbits_{\tilde{\Gamma}_G}(x,y) = \tilde{Z}_G\left(f_X(x),\ldots,f_X(x^k),f_Y(y),\ldots,f_Y(y^k)\right). \tag{2.12}$$

2.4. The Dihedral Group, its Cycle Index and its Extension

To apply these results to the problem of counting distinct DNA necklaces, we will need to describe the relevant group action and compute its (bipartite) cycle index. The set of elements acted on by the group is given by the α containers in the DNA necklace and this set can be partitioned into two groups: containers of black beads and containers of white beads. We consider two DNA necklaces to be the same if one can be *rotated* or *reflected* into the other. These symmetries can be described by an action of the dihedral group, which we will denote by D_{2M}, where we have $\alpha = 2M$. The rotational and reflective symmetries are what distinguishes the case of periodic DNA chains from linear, fixed boundary condition chains studied in [31] and elsewhere.

A fundamental fact about D_{2M} is that it is generated by two elements r and s, where r is a reflection satisfying $r^2 = 1$, and s is a rotation of order M. Therefore, to describe a group action of D_{2M} on a DNA necklace it suffices to give the action of r and s. In Fig. 4 the action of such a rotation on the necklace is illustrated, while in Figs. 5 and 6 the action of a reflection is illustrated for the cases where M is odd and even, respectively. It is clear that the resulting group action is partition-preserving.

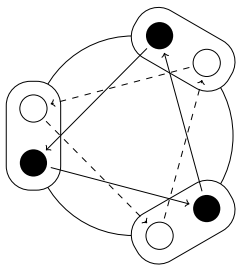

Fig. 4. The action of a rotation $s \in D_{2M}$ on the DNA necklace.

To compute the bipartite cycle index of this group action, we will treat reflections and rotations separately. To begin with, we can see from Fig. 4 that rotations act symmetrically on the black and white containers in the DNA necklace. Thus, the terms of the cycle index polynomial corresponding to rotations will be symmetric in the x_i and y_i. The natural action of the cyclic group C_M on the M containers in a partition is given by [25]:

$$Z_{C_M}(x_1,\ldots,x_M) = \sum_{d|M} \varphi(d) x_d^{M/d}, \tag{2.13}$$

where $\varphi(d)$ is defined to be the number of natural numbers less than d that are coprime to it (the *Euler totient function*). Note that 1 is considered to be coprime to all natural numbers, and so in particular $\varphi(d) > 0$. Exactly half of the elements of D_{2M} are rotations, and thus the rotational part of the bipartite cycle index $\tilde{Z}_{D_{2M}}$ is given by $\frac{1}{2}\sum_{d|M}\varphi(d)x_d^{M/d}y_d^{M/d}$.

The reflective part of the group D_{2M}, on the other hand, acts differently depending on the parity of M. Suppose first that M is odd, in which case a typical reflection is illustrated in Fig. 5. Each of the M possible reflections occur across an axis consisting of one black container and one white container, both of which are fixed by the reflection. The rest of the containers are split into 2-cycles, and thus the bipartite cycle index $\tilde{Z}_{D_{2M}}$ for odd M is given by

$$\tilde{Z}_{D_{2M}}(x_1,\ldots,x_M,y_1,\ldots,y_M) = \frac{1}{2}\sum_{d|M}\varphi(d)x_d^{M/d}y_d^{M/d} + \frac{1}{2}x_1y_1x_2^{(M-1)/2}y_2^{(M-1)/2}. \tag{2.14}$$

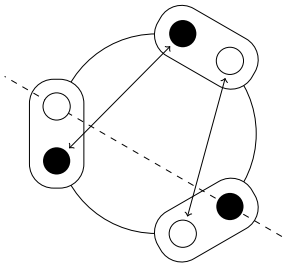

Fig. 5. The action of a reflection $r \in D_{2M}$ on the DNA necklace for the case where M is odd.

If M is even, a typical reflection is illustrated in Fig. 6. In this case, each possible reflection occurs across an axis consisting of either two white containers or two black containers. The rest of the containers again split into 2-cycles. Thus, the bipartite cycle index $\tilde{Z}_{D_{2M}}$ for even M is given by

$$\tilde{Z}_{D_{2M}}(x_1,\ldots,x_M,y_1,\ldots,y_M) = \frac{1}{2}\sum_{d|M}\varphi(d)x_d^{M/d}y_d^{M/d} + \frac{1}{4}x_1^2x_2^{(M-2)/2}y_2^{M/2} + \frac{1}{4}y_1^2y_2^{(M-2)/2}x_2^{M/2}. \tag{2.15}$$

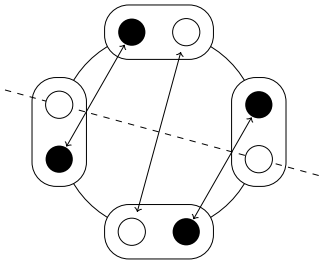

Fig. 6. The action of a reflection $r \in D_{2M}$ on the DNA necklace, for the case where M is even.

2.5. Generating Functions as Formal Power Series

In our particular application of Pólya theory, the *elements* we are coloring are the α containers in the DNA necklace and the *color* of a particular container is defined to be the number of black or white beads it contains. As each container must contain at least one bead, the set of colors is given by $\mathbb{N}^+$. We are interested in the total number of black and white beads, so the weight of each color will be given quite simply by $\omega(n) = n$ for each $n \in \mathbb{N}^+$. This weighting corresponds to the generating function (2.5) $f_\omega(x) = x + x^2 + x^3 + \ldots$.

To compute the number of distinct DNA necklaces with N_{AT} white beads and N_{GC} black beads, we need to calculate the coefficient of $x^{N_{AT}}y^{N_{GC}}$ in (2.12), where the bivariate cycle index is given by the appropriate $\tilde{Z}(D_{2M})$ from Section 2.4 and the weight generating function is given by $f_\omega(x)$. This requires us to calculate the coefficients of specific terms in $f_\omega(x)^n = (x + x^2 + x^3 + \ldots)^n$ for potentially large n. However, doing this expansion naively requires many computing steps, whose number grows exponentially fast as n increases. Thus, this approach is impractical. Fortunately, there exists a way to bypass this problem: treating $f_\omega(x)$ as a formal power series, we can manipulate it into a form that makes such computations significantly faster.

An introduction to the theory of formal power series can be found, for instance, in [42]. For our purposes, we will only need the fact that a form of the binomial theorem holds in this setting:

Lemma 3. *Letting* $(1-x)^{-n}$ *denote the formal inverse of* $(1-x)^n$*, we have*

$$(1-x)^{-n} = \sum_{k=0}^{\infty}\binom{n+k-1}{n-1}x^k. \tag{2.16}$$

This implies the following useful lemma regarding powers of $f_\omega(x)$:

Lemma 4. *As a formal power series $f_\omega(x)^n$ can be written as $f_\omega(x)^n = \sum_{k=0}^{\infty} \binom{n+k-1}{n-1} x^{n+k}$.*

Proof. Note that $xf_\omega(x) = x^2 + x^3 + \ldots = f_\omega(x) - x$. Rearranging this for $f_\omega(x)$, we see that $f_\omega(x) = x(1-x)^{-1}$, and hence $f_\omega(x)^n = x^n(1-x)^{-n}$. The result now follows from Lemma 3. □

In contrast to naively expanding powers of $f_\omega(x)$, computing binomial coefficients is computationally inexpensive, taking at most a linear number of steps in n.

We now list a few results that will come in handy later, when we describe an explicit algorithm for computing the number of distinct DNA necklaces with the given constraints.

Lemma 5. *The coefficient of x^r in $f_\omega(x^a)^b$ is given by*

$$\left[f_\omega(x^a)^b\right]_r = \begin{cases} 1 & \text{if } b = 0 \text{ and } a = 0 \\ 0 & \text{if } b = 0 \text{ and } a > 0 \\ 0 & \text{if } b > 0 \text{ and } a \nmid r \text{ or } r < ab \\ \binom{r/a-1}{b-1} & \text{otherwise.} \end{cases} \tag{2.17}$$

Lemma 6. *The coefficient of x^r in $f_\omega(x^{a_1})^{b_1} \cdot f_\omega(x^{a_2})^{b_2}$ is given by*

$$\left[f_\omega(x^{a_1})^{b_1} \cdot f_\omega(x^{a_2})^{b_2}\right]_r = \sum_{k=0}^{r} \left[f_\omega(x^{a_1})^{b_1}\right]_k \left[f_\omega(x^{a_2})^{b_2}\right]_{r-k}. \tag{2.18}$$

3. THE ALGORITHM FOR COMPUTING THE NUMBER OF DISTINCT VALID NECKLACES

Now we are able to evaluate the number of distinct necklaces, which correspond to a particular value of alternations α. The algorithm is fairly straightforward and efficient. Its implementation requires the following steps:

a) Set constraint parameters, N_{AT}, N_{GC}, and $\alpha = 2M$.

b) Choose partitioned cycle index polynomial of the Dihedral group based on parity of M. If M is odd, use (2.14), while for M even use (2.15).

c) By the corollary to Pólya's Enumeration Theorem (2.12), we know that the number of necklaces, up to symmetry, is given by

$$\mathrm{Orbits}_{\tilde{\Gamma}_G}(x,y) = \tilde{Z}_G\left(f_X(x), \ldots, f_X(x^k), f_Y(y), \ldots, f_Y(y^k)\right). \tag{3.1}$$

If M is odd, using the outcome of the previous step we get

$$\begin{aligned} \mathrm{Orbits}_{\tilde{\Gamma}_G}(x,y) = {} & \frac{1}{2M}\sum_{d|M} \varphi(d) f^{M/d}(x^d) f^{M/d}(y^d) \\ & + \frac{1}{2} f(x) f(y) f^{(M-1)/2}(x^2) f^{(M-1)/2}(y^2). \end{aligned} \tag{3.2}$$

If M is even, then we have

$$\begin{aligned} \mathrm{Orbits}_{\tilde{\Gamma}_G}(x,y) = {} & \frac{1}{2M}\sum_{d|M} \varphi(d) f^{M/d}(x^d) f^{M/d}(y^d) \\ & + \frac{1}{4} f^2(x) f^{(M-2)/2}(x^2) f^{M/2}(y^2) + \frac{1}{4} f^2(y) f^{(M-2)/2}(y^2) f^{M/2}(x^2). \end{aligned} \tag{3.3}$$

d) Every term in the polynomial produced by (3.1) will be of the form in (2.17) or (2.18). The number of necklaces with N_{AT} white beads and N_{GC} black beads is given by the coefficient of the term $x^{N_{AT}}y^{N_{GC}}$. To calculate the total number of necklaces, simply sum over each of these terms appearing in the polynomial.

A Python computer code implementing this algorithm is presented in the Appendix.

In order to illustrate the application of this algorithm let us consider a simple, but not trivial case: We set $\alpha = 2M = 10$, $N_{\mathrm{AT}} = 8$, $N_{\mathrm{GC}} = 6$. Clearly, $M = 5$ is odd, so identifying white beads with AT base pairs and black beads with GC base pairs, we have the cycle index

$$\begin{aligned}\tilde{Z}(\tilde{D}_{10}) &= \frac{1}{2}Z(\tilde{C}_5) + \frac{1}{2}x_1y_1(x_2)^2(y_2)^2 \\ &= \frac{1}{5}\sum_{d|5}\varphi(d)(x_d)^{5/d}(y_d)^{5/d} + \frac{1}{2}x_1y_1(x_2)^2(y_2)^2. \end{aligned} \tag{3.4}$$

Now the partitioned Pólya Enumeration Theorem tells us that we can put the generating functions $f_W\left(x^d\right)$ and $f_B\left(y^d\right)$ in place of the x_d and y_d, respectively, to find the generating function of fixed orbits. So we have

$$\begin{aligned}\mathrm{Orbits}_{\tilde{\Gamma}_G}(x,y) =& \frac{1}{2\cdot 5}\left[1(x+x^2+x^3+\dots)^5(y+y^2+y^3+\dots)^5\right. \\ &\left. + 4(x^5+x^{10}+x^{15}+\dots)(y^5+y^{10}+y^{15}+\dots)\right] \\ &+\frac{1}{2}(x+x^2+\dots)(x^2+x^4+\dots)^2(y+y^2+\dots)(y^2+y^4+\dots)^2. \end{aligned} \tag{3.5}$$

Let us first look at the cyclic part. Since 5 is prime, the only two integers that divide it are 1 and 5, so this polynomial will be

$$\frac{1}{2\cdot 5}\left[1(x+x^2+x^3+\dots)^5(y+y^2+y^3+\dots)^5 + 4(x^5+x^{10}+x^{15}+\dots)(y^5+y^{10}+y^{15}+\dots)\right].$$

Now we try to extract the coefficients of terms that are allowed. These are the terms in $x^{N_{\mathrm{AT}}}$ and $y^{N_{\mathrm{GC}}}$ and we can use (2.17) in order to calculate these coefficients directly. In this case, there will be no contribution from the second term, as there are no terms in x^8 and y^6. So the total cyclic contribution will be (with $r = 8$ and $r = 6$ for the respective cases and $a = 1,\ b = 5$ for both)

$$\frac{1}{10}\binom{N_{\mathrm{GC}}-1}{5-1}\binom{N_{\mathrm{AT}}-1}{5-1} = \frac{1}{10}\binom{5}{4}\binom{7}{4} = \frac{175}{10}.$$

Then the same coefficient identifying process can be followed for the reflective part. Now the polynomial is given by

$$\frac{1}{2}(x+x^2+\dots)(x^2+x^4+\dots)^2(y+y^2+\dots)(y^2+y^4+\dots)^2.$$

So for both x and y the coefficients will come from the product of two series, one of them squared. Thus, the relevant terms will come in a series of products given in (2.18). In y the sum of coefficients contracts to a single element. That contribution is simply $\binom{1}{0}\binom{1}{1} = 1$. In x, however, there will be terms from $x^2\cdot x^6$ as well as $x^4\cdot x^4$. So then, the sum will be

$$\binom{1}{0}\binom{3}{1} + \binom{3}{0}\binom{1}{1} = 4,$$

giving a total contribution of $\frac{1}{2}(1+4) + \frac{175}{10} = 20$. Thus, there are 20 DNA chains with 8 AT base pairs, 6 GC base pairs and 10 alternations.

4. NUMERICAL RESULTS

The developed algorithm for calculating the number of distinct DNA chains having α alternations can be used to produce the pdf of α, $P(\alpha)$, which afterwards can be compared to pdfs numerically obtained from Monte Carlo (MC) simulations. In Figs. 7a and 7b we present such pdfs for a DNA chain containing $N = 100$ base pairs. In particular, we consider the case of $N_{AT} = 40$, $N_{GC} = 60$ in Fig. 7a and the case of $N_{AT} = 50$, $N_{GC} = 50$ in Fig. 7b. From Figs. 7a and 7b we clearly see that the results obtained by the algorithm presented in Section 3 (empty circles) and by MC simulations of DNA chains with $N = 100$ base pairs (filled stars) agree very well. The slight differences between them are to be expected, as the number of possible chains is generally very large. For instance, in the case of $N_{AT} = 50$, $N_{GC} = 50$ and $\alpha = 50$, the number of possible DNA chains is of the order of 10^{25} possible necklaces. Thus, in general, the number of performed MC simulations cannot get close to the actual total number of possible chains. Nevertheless, although the results of Figs. 7a and 7b were obtained by only $N_{MC} = 20000$ MC simulations, they manage to capture the theoretically obtained pdf quite accurately. Of course, it is expected that increasing the number of MC simulations will improve the accuracy of the numerical results. As a measure of this accuracy we can consider the total absolute difference

$$d(N_{MC}) = \sum_{\alpha} |P_{MC}(N_{MC}, \alpha) - P(\alpha)|, \tag{4.1}$$

between the two distributions. In (4.1) $P_{MC}(N_{MC}, \alpha)$ is the probability of α alternations obtained by N_{MC} MC simulations, $P(\alpha)$ is the one obtained theoretically, while the sum is performed over

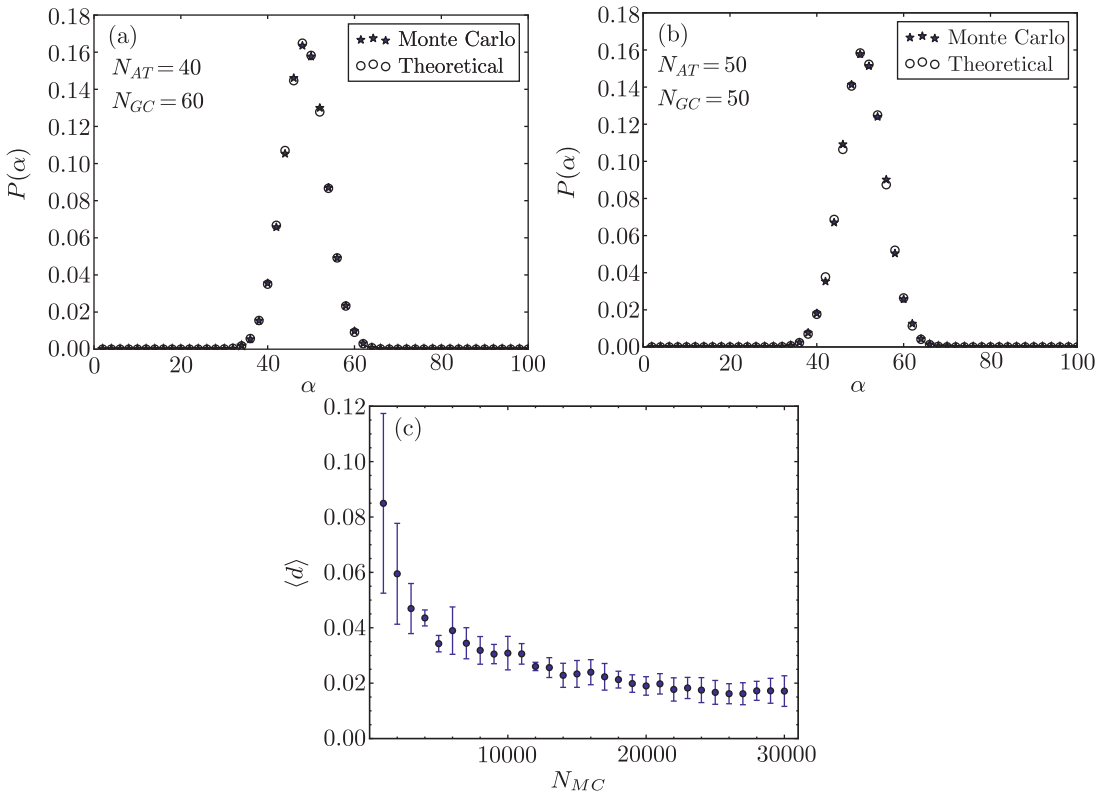

Fig. 7. Comparison of the pdf $P(\alpha)$ of the number of alternations α, obtained by the algorithm presented in Section 3 [empty circles in panels (a) and (b)] and by randomly created DNA chains of $N = 100$ base pairs through MC simulations [filled stars in panels (a) and (b)]. The pdfs for $N_{AT} = 40$, $N_{GC} = 60$ and $N_{AT} = 50$, $N_{GC} = 50$ are presented in panels (a) and (b), respectively. The number of MC simulations used in (a) and (b) are $N_{MC} = 20000$. (c) The evolution of the average total absolute difference $\langle d \rangle$ between the theoretically and the numerically obtained pdfs as a function of N_{MC} for the case of $N_{AT} = 50$, $N_{GC} = 50$. The values of $\langle d \rangle$ are obtained as the average of the quantity (4.1) evaluated for 5 different sets of N_{MC} runs. The error bars denote the corresponding standard deviations.

all possible values of α. From the results of Fig. 7c where we plot the averaged value of $d(N_{MC})$ over 5 sets of N_{MC} MC simulations as a function of N_{MC} we see that, as the number of simulations increases, the numerical results get closer to the theoretical ones.

The results of Fig. 7c clearly show that, in order to study the dynamical properties of DNA chains, statistical analysis performed over a few thousands of MC generated random chains (even of the order of 5000) would suffice, as such numbers of MC simulations are enough for capturing quite accurately the influence of alternations on the system's dynamics.

The shape of the pdfs in Figs. 7a and 7b suggests that they could possibly be fitted by Gaussian distributions. This is actually true as we can see from the results of Fig. 8, where we performed such a fit for the theoretically obtained pdf of Fig. 7b. The Gaussian approximation of the pdfs has several advantages as it allows us to easily quantify the influence of different variables on the number of alternations. Let us first look at the effect of increasing the number of only one type of base pair, keeping constant the number of the other type of base pair. In Fig. 9 we present some pdfs of α for $N_{AT} = 100$ and increasing values of N_{GC} from 25 up to 2500. Starting from small values of N_{GC}, we find a very "lopsided" and narrow distribution which, as N_{GC} increases, becomes gradually more symmetric and spreads out, up to a value of $N_{GC} = 200$ (see Fig. 10b below). Then, increasing N_{GC} further, as the numbers of different types of base pairs become more dissimilar, we again find gradually more unbalanced pdfs with sharp peaks. The very "lopsided" distributions are obtained when the minority base pairs are significantly less than the majority ones and therefore are spread out and isolated among the others. In this case the distribution is sharply peaked around the corresponding maximum possible number of alternations. For the $N_{AT} = 100$, $N_{GC} = 25$ case this number is $\alpha = 50$, while for the $N_{AT} = 100$, $N_{GC} = 2500$ case it is $\alpha = 200$.

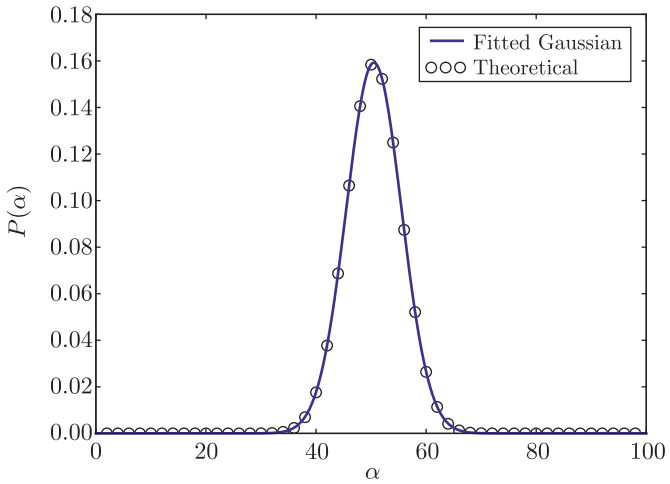

Fig. 8. Fitting by a Gaussian of the theoretical pdf of Fig. 7b (empty circles) with $N_{AT} = 50$, $N_{GC} = 50$. The mean of the Gaussian is $\alpha_0 = 50.5$ and standard deviation $\sigma_\alpha = 5.1$.

These changes of the distributions are quantitatively presented in Fig. 10 through the variations of the fitted Gaussian characteristics. The increase of the mean value α_0 of the Gaussian fits as the number N_{GC} increases is shown in Fig. 10a. The upper limit of α_0 is 200, when N_{GC} becomes much larger than N_{AT}. The dependence of the width (standard deviation) σ_α of the Gaussian fits on N_{GC} is depicted in Fig. 10b. The initial increase with N_{GC} corresponds to the spreading out of the distributions when the numbers of base pairs become more similar. Further increase of the N_{GC} values pushes the pdfs to the other extreme and the lopsidedness comes through again, resulting in narrower distributions (see Fig. 9). This results in the decrease of σ_α for large values of N_{GC}. Finally, in Fig. 10c we observe that, as N_{GC} increases, the maximum probability of the pdfs initially decreases rapidly and then increases slowly, in accordance with the results of Fig. 9 and, of course, with the fact that it is inversely proportional to the standard deviation of the Gaussian fit.

Let us now focus our attention on the effect of the increment of the total number of base pairs $N = N_{AT} + N_{GC}$, i. e., the total "length" of the DNA chain, when the ratio $N_{GC} : N_{AT}$ is kept constant. Such cases are presented in Fig. 11, where we plot several pdfs for different values of N but for fixed ratios $N_{GC} : N_{AT}$. In particular, the values of the ratios $N_{GC} : N_{AT}$ are $1 : 1$ in

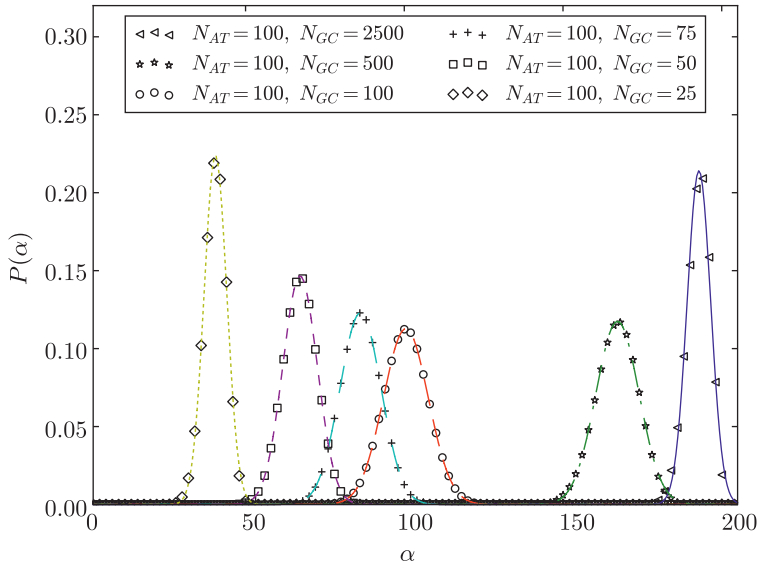

Fig. 9. Pdfs of α for fixed number of AT base pairs ($N_{AT} = 100$) and increasing values of N_{GC}. Points correspond to the theoretically obtained values of the pdfs, while curves correspond to the Gaussian fits of these points. Note that even for long DNA chains the value of α cannot exceed $\alpha = 200$.

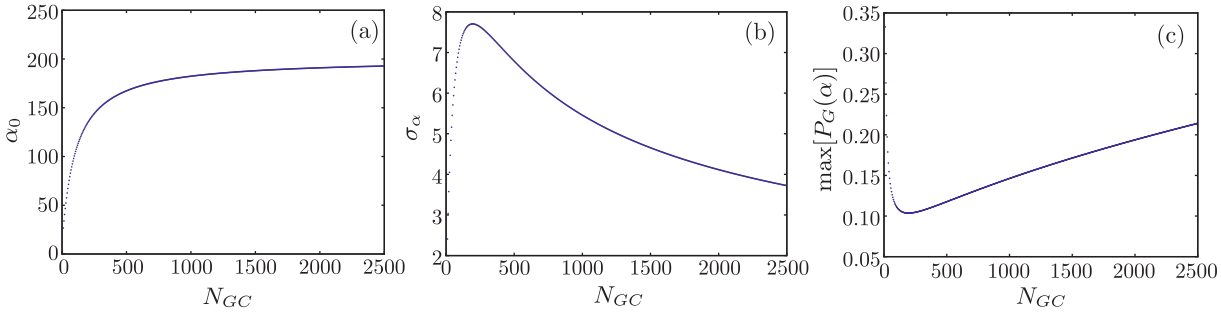

Fig. 10. The effect of increasing the number N_{GC} of the GC base pairs for a fixed number of AT base pairs ($N_{AT} = 100$) on the Gaussian fit $P_G(\alpha)$ of the pdf values of α, and in particular on (a) the mean value α_0, (b) the standard deviation σ_α and (c) the maximum probability $\max[P_G(\alpha)]$. Some of these pdfs are shown in Fig. 9.

panel (a), 2 : 1 in (b) and 6 : 1 in (c). In all cases the pdfs are fitted by appropriate Gaussian distributions whose characteristics are plotted in Fig. 12 as a function of N. From the results of Figs. 11 and 12 we see that, as the total number N of base pairs increases, the pdfs become broader, and consequently their maximum value decreases. This means that for large N more α values have a relatively high probability to appear in a randomly created DNA chain. In addition, increasing the ratio $N_{GC} : N_{AT}$ results in a decrease of the spreading, as evidenced by the lower standard deviation in Fig. 12b and the higher maximum probability in Fig. 12c. A linear relationship between N and the mean α_0 is observed for all ratios, with the slope of the line influenced by the ratio. The slope m for each case is: $m = 0.25$ for ratio 6 : 1, $m = 0.45$ for 2 : 1 and $m = 0.5$ for 1 : 1.

5. CONCLUSIONS

Motivated by the possibility that the number α of base pair alternations in a circular or periodic DNA chain might affect the dynamics of the system, we have found a probability distribution for this number. Algorithms for such distributions are known for linear DNA sequences with fixed boundary conditions [31]. The introduction of the periodic boundary conditions we consider in our study makes the counting of alternations a much more complicated problem due to the appearance of additional rotational and reflectional symmetries. To account for the additional complexity arising from these symmetries, we have implemented Pólya counting theory. In particular, extending

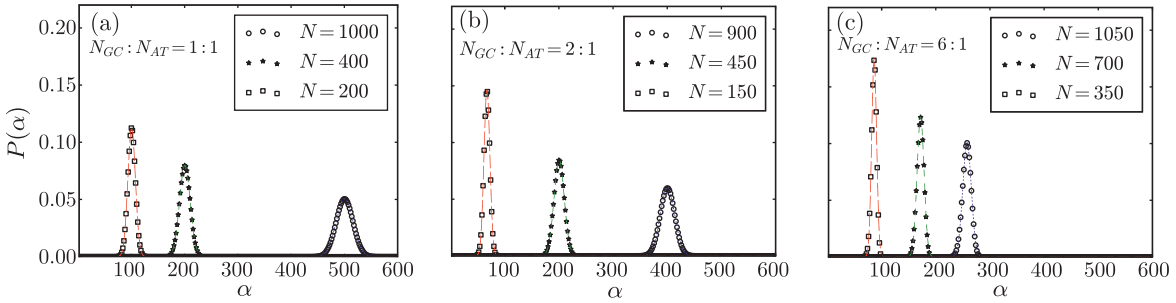

Fig. 11. Pdfs of α for fixed ratios $N_{GC} : N_{AT} = 1 : 1$ (a), $2 : 1$ (b) and $6 : 1$ (c). Points correspond to the theoretically obtained values of the pdfs, while curves correspond to the Gaussian fits of these points.

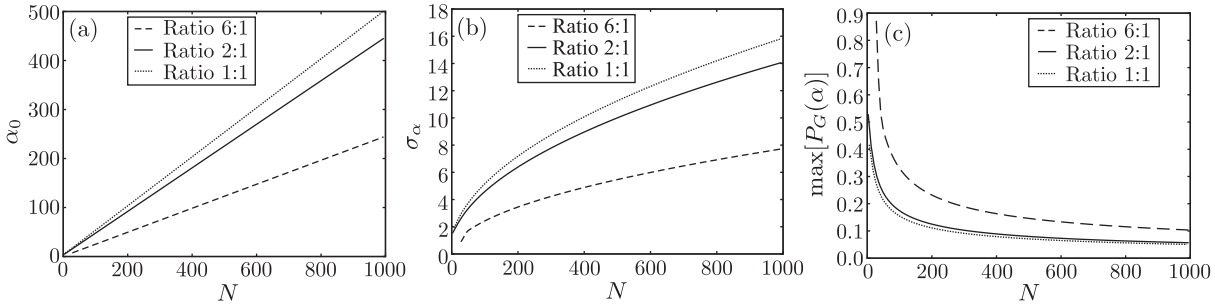

Fig. 12. The effect of increasing the total number of base pairs N for fixed ratios $N_{GC} : N_{AT}$ on the parameters of the Gaussian fit $P_G(\alpha)$ of the pdf for α: (a) the mean value α_0, (b) the standard deviation σ_α and (c) the maximum probability $\max[P_G(\alpha)]$. Some of these pdfs are shown in Fig. 11.

Pólya's Enumeration Theorem for a partition-preserving group action on a partitioned set, we have constructed a well-defined algorithm for calculating the number of DNA chains having a given number of alternations for particular values of the number of AT (N_{AT}) and GC (N_{GC}) base pairs.

The obtained theoretical results were compared with numerically constructed pdfs through MC simulations. We found that, in general, creating a few thousands of random DNA chains (around 5000) by MC simulations we can approximate quite accurately the theoretical pdf of α. This means that a statistical analysis of these DNA chains will suffice to uncover the potential influence of heterogeneity on the dynamic behavior of the considered DNA model.

In addition, approximating the obtained pdfs by Gaussians, we investigated the effect of the number of the two base pairs, as well as their ratio on various characteristics of the pdfs, like their mean value, their standard deviation and their maximum.

APPENDIX

Here we present a Python computer code implementing the algorithm of Section 3. The function `necklace_count(n, B, W)` returns the total number of possible necklaces under the symmetry constraints with $2n$ alternations, B black beads and W white beads.

```
from math import gcd
# Compute binomial coefficients in linear time.
def binomial (n, k):
    if k > n or k < 0:
        return 0
    if k == 0:
        return 1
    if k > n//2:
        return binomial(n, n-k)
    return (n * binomial(n-1, k-1)) // k

# Compute the Euler totient function \phi(n), which
```

```
# gives the number of integers 0 < d <= n that are
# relatively prime to n.
def totient(n):
    count = 0
    for d in range(1, n+1):
        if gcd(d, n) == 1:
            count += 1
    return count

# Get the x^r coefficient of our weight generating functions f(x^m)^n,
# where:
#    f(x) = x + x^2 + x^3 + ...
def weight_gf(r, m, n):
    if n == 0:
        if r == 0:
            return 1
        return 0
    if r%m != 0:
        return 0
    if (r//m) < n:
        return 0
    return binomial((r // m)-1, n-1)

# Get the x^r coefficient of a binary product of weight generating
# functions f(x^m1)^n1 * f(x^m2)^n2, where:
#    f(x) = x + x^2 + x^3 + ...
def binary_weight_gf(r, m1, n1, m2, n2):
    total = 0
    for i in range(1, r):
        total += weight_gf(i, m1, n1) * weight_gf(r-i, m2, n2)
    return total

# Compute the number of necklaces up to dihedral symmetry with
# 2n alternations, B black beads and W white beads.
def necklace_count(n, B, W):
    # First we count the contributions from the cyclic part
    # of the cycle index.
    count = 0
    for d in range(1, n+1):
        if n%d != 0:
            continue
        count += totient(d) * weight_gf(B, d, n//d) *
                 weight_gf(W, d, n//d)
    # Next we count the contributions from the dihedral part
    # of the cycle index.
    if n%2 == 0:
        count += (weight_gf(B, 2, n//2) *
                  binary_weight_gf (W, 1, 2, 2, (n-2)//2) * (n//2))
        count += (weight_gf(W, 2, n//2) *
                  binary_weight_gf(B, 1, 2, 2, (n-2)//2) * (n//2))
    else:
        count += (binary_weight_gf(B, 1, 1, 2, (n-1)//2) *
                  binary_weight_gf(W, 1, 1, 2, (n-1)//2) * n)
    return count // (2*n)
```

ACKNOWLEDGMENTS

M. H. and G. P-J. acknowledge financial assistance from the National Research Foundation (NRF) of South Africa towards this research. G.K. and Ch.S. were supported by the Erasmus+/ International Credit Mobility KA107 program. Ch.S. acknowledges support by the NRF of South Africa (IFRR and CPRR Programmes), the UCT (URC Conference Travel Grant) and thanks Hans-Peter Kunzi for useful discussions.

REFERENCES

1. Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P., *Essential Cell Biology,* 3rd ed., New York: Garland Sci., 2009.
2. Alexandrov, B. S., Gelev, V., Monisova, Y., Alexandrov, L. B., Bishop, A. R., Rasmussen, K. Ø., and Usheva, A., A Nonlinear Dynamic Model of DNA with a Sequence-Dependent Stacking Term, *Nucleic Acids Res.,* 2009, vol. 37, no. 7, pp. 2405–2410.
3. Alexandrov, B. S., Gelev, V., Yoo, S. W., Bishop, A. R., Rasmussen, K. Ø., and Usheva, A., Toward a Detailed Description of the Thermally Induced Dynamics of the Core Promoter, *PLoS Comput. Biol.,* 2009, vol. 5, no. 3, e1000313, 10 pp.
4. Alexandrov, A. S., Gelev, V., Yoo, S. W., Alexandrov, L. B., Fukuyo, Y., Bishop, A. R., Rasmussen, K. Ø., and Usheva, A., DNA Dynamics Play a Role as a Basal Transcription Factor in the Positioning and Regulation of Gene Transcription Initiation, *Nucleic Acids Res.,* 2010, vol. 38, no. 6, pp. 1790–1795.
5. Apostolaki, A. and Kalosakas, G., Targets of DNA-Binding Proteins in Bacterial Promoter Regions Present Enhanced Probabilities for Spontaneous Thermal Openings, *Phys. Biol.,* 2011, vol. 8, no. 2, 026006, 31 pp.
6. Ares, S., Voulgarakis, N. K., Rasmussen, K. Ø., and Bishop, A. R., Bubble Nucleation and Cooperativity in DNA Melting, *Phys. Rev. Lett.,* 2005, vol. 94, no. 3, 035504, 4 pp.
7. Ares, S. and Kalosakas, G., Distribution of Bubble Lengths in DNA, *Nano Lett.,* 2007, vol. 7, no. 2, pp. 307–311.
8. Brualdi, R. A., Pólya Counting, in *Introductory Combinatorics,* 5th ed., Upper Saddle River, N.J.: Prentice Hall, 2010, pp. 541–581.
9. Burnside, W., *Theory of Groups of Finite Order,* Cambridge: Cambridge Univ. Press, 1897.
10. Chetverikov, A. P., Ebeling, W., Lakhno, V. D., Shigaev, A. S., and Velarde, M. G., On the Possibility That Local Mechanical Forcing Permits Directionally-Controlled Long-Range Electron Transfer along DNA-Like Molecular Wires with No Need of an External Electric Field, *Eur. Phys. J. B,* 2016, vol. 89, no. 4, Art. 101, 10 pp.
11. Choi, Ch. H., Kalosakas, G., Rasmussen, K. Ø., Hiromura, M., Bishop, A. R., and Usheva, A., DNA Dynamically Directs Its Own Transcription Initiation, *Nucleic Acids Res.,* 2004, vol. 32, no. 4, pp. 1584–1590.
12. Choi, Ch. H., Rapti, Z., Gelev, V., Hacker, M. R., Alexandrov, B. S., Park, E. J., Park, J. S., Horikoshi, N., Smerzi, A., Rasmussen, K. Ø., Bishop, A. R., and Usheva, A., Profiling the Thermodynamic Softness of Adenoviral Promoters, *Biophys. J.,* 2008, vol. 95, no. 2, pp. 597–608.
13. Dauxois, T., Peyrard, M., and Bishop, A. R., Dynamics and Thermodynamics of a Nonlinear Model for DNA Denaturation, *Phys. Rev. E,* 1993, vol. 47, no. 1, pp. 684–695.
14. Hennig, D., Control of Electron Transfer in Disordered DNA under the Impact of Viscous Damping and an External Periodic Field, *Eur. Phys. J. B,* 2002, vol. 30, no. 2, pp. 211–218.
15. Herstein, I. N., *Abstract Algebra,* 3rd ed., New York: Wiley, 1999.
16. Huang, H.-H. and Lindblad, P., Wide-Dynamic-Range Promoters Engineered for Cyanobacteria, *J. Biol. Eng.,* 2013, vol. 7, no. 1, Art. 10, 11 pp.
17. Kalosakas, G., Charge Transport in DNA: Dependence of Diffusion Coefficient on Temperature and Electron-Phonon Coupling Constant, *Phys. Rev. E,* 2011, vol. 84, no. 5, 051905, 6 pp.
18. Kalosakas, G. and Ares, S., Dependence on Temperature and Guanine-Cytosine Content of Bubble Length Distributions in DNA, *J. Chem. Phys.,* 2009, vol. 130, no. 23, 235104, 7 pp.
19. Kalosakas, G., Ngai, K. L., and Flach, S., Breather-Induced Anomalous Charge Diffusion, *Phys. Rev. E,* 2005, vol. 71, no. 6, 061901, 7 pp.
20. Kalosakas, G., Rasmussen, K. Ø., Bishop, A. R., Choi, Ch. H., and Usheva, A., Sequence-Specific Thermal Fluctuations Identify Start Sites for DNA Transcription, *Europhys. Lett.,* 2004, vol. 68, no. 1, pp. 127–133.
21. Kalosakas, G., Rasmussen, K. Ø., and Bishop, A. R., Non-Exponential Decay of Base-Pair Opening Fluctuations in DNA, *Chem. Phys. Lett.,* 2006, vol. 432, nos. 1–3, pp. 291–295.
22. Kolpakov, R., Bana, G., and Kucherov, G., mreps: Efficient and Flexible Detection of Tandem Repeats in DNA, *Nucleic Acids Res.,* 2003, vol. 31, no. 13, pp. 3672–3678.
23. Lewin, B., *Genes VIII,* 8th ed., Upper Saddle River, N.J.: Pearson/Prentice Hall, 2004.

24. Li, W., The Study of Correlation Structures of DNA Sequences: A Critical Review, *Comput. Chem.,* 1997, vol. 21, no. 4, pp. 257–271.
25. van Lint, J. H. and Wilson, R. M., Pólya Theory of Counting, in *A Course in Combinatorics,* Cambridge: Cambridge Univ. Press, 1992, pp. 522–535.
26. Nowak-Lovato, K., Alexandrov, L. B., Banisadr, A., Bauer, A. L., Bishop, A. R., Usheva, A., Mu, F., Hong-Geller, E., Rasmussen, K. Ø., Hlavacek, W. S., and Alexandrov, B. S., Binding of Nucleoid-Associated Protein Fis to DNA Is Regulated by DNA Breathing Dynamics, *PLoS Comput. Biol.,* 2013, vol. 9, no. 1, e1002881.
27. Peyrard, M., Nonlinear Dynamics and Statistical Physics of DNA, *Nonlinearity,* 2004, vol. 17, no. 2, R1–R40.
28. Peyrard, M. and Farago, J., Nonlinear Localization in Thermalized Lattices: Application to DNA, *Phys. A,* 2000, vol. 288, nos. 1–4, pp. 199–217.
29. Pólya, G. and Read, R. C., Chemical Compounds, in *Combinatorial Enumeration of Groups, Graphs, and Chemical Compounds,* New York: Springer, 1987, pp. 58–74.
30. Régnier, M., A Unified Approach to Word Occurrence Probabilities. Combinatorial Molecular Biology, *Discrete Appl. Math.,* 2000, vol. 104, nos. 1–3, pp. 259–280.
31. Robin, S. and Daudin, J. J. Exact Distribution of Word Occurrences in a Random Sequence of Letters, *J. Appl. Probab.,* 1999, vol. 36, no. 1, pp. 179–193.
32. Robin, S. and Schbath, S., Numerical Comparison of Several Approximations of the Word Count Distribution in Random Sequences, *J. Comput. Biol.,* 2001, vol. 8, no. 4, pp. 349–359.
33. Tabi, C. B., Dang Koko, A., Oumarou Doko, R., Ekobena Fouda, H. P., and Kofané, T. C., Modulated Charge Patterns and Noise Effect in a Twisted DNA Model with Solvent Interaction, *Phys. A,* 2016, vol. 442, pp. 498–509.
34. Schbath, S., Compound Poisson Approximation of Word Counts in DNA Sequences, *ESAIM: Probab Statist.,* 1995, vol. 1, pp. 1–16.
35. Schbath, S., Prum, B., and de Turckheim, E., Exceptional Motifs in Different Markov Chain Models for a Statistical Analysis of DNA Sequences, *J. Comput. Biol.,* 1995, vol. 2, pp. 417–437.
36. Skokos, Ch., Hillebrand, M., Schwellnus, A., and Kalosakas, G., in preparation (2018).
37. Tapia-Rojo, R., Mazo, J. J., and Falo, F., Thermal and Mechanical Properties of a DNA Model with Solvation Barrier, *Phys. Rev. E,* 2010, vol. 82, no. 3, 031916, 8 pp.
38. Tapia-Rojo, R., Mazo, J. J., Hernández, J. A., Peleato, M. L., Fillat, M. F., and Falo, F., Mesoscopic Model and Free Energy Landscape for Protein-DNA Binding Sites: Analysis of Cyanobacterial Promoters, *PLoS Comput. Biol.,* 2014, vol. 10, no. 10, e1003835.
39. Theodorakopoulos, N., DNA Denaturation Bubbles at Criticality, *Phys. Rev. E,* 2008, vol 77, no. 3, 031919, 8 pp.
40. Voulgarakis, N. K., Kalosakas, G., Rasmussen, K. Ø., and Bishop, A. R., Temperature-Dependent Signatures of Coherent Vibrational Openings in DNA, *Nano Lett.,* 2004, vol. 4, no. 4, pp. 629–632.
41. Yakushevich, L. V., *Nonlinear Physics of DNA,* 2nd ed., New York: Wiley-VCH, 2004.
42. Zariski, O. and Samuel, P., Polynomial and Power Series Rings, in *Commutative Algebra: Vol. 2,* Grad. Texts in Math., vol. 29, New York: Springer, 1975, pp. 129–247.
43. Zoli, M., Anharmonic Stacking in Supercoiled DNA, *J. Phys. Condens. Matter,* 2012, vol. 24, no. 19, 195103, 23 pp.
44. Zoli, M., Twist versus Nonlinear Stacking in Short DNA Molecules, *J. Theor. Biol.,* 2014, vol. 354, pp. 95–104.