Article

DNA Sequence and Structure under the Prism of Group Theory and Algebraic Surfaces

Michel Planat \(^1\,*\,+\), Marcelo M. Amaral \(^2\,+\), Fang Fang \(^2\), David Chester \(^2\), Raymond Aschheim \(^2\) and Klee Irwin \(^2\)

\(^1\) Institut FEMTO-ST CNRS UMR 6174, Université de Bourgogne-Franche-Comté, F-25044 Besançon, France
\(^2\) Quantum Gravity Research, Los Angeles, CA 90290, USA
* Correspondence: michel.planat@femto-st.fr
† These authors contributed equally to this work.

Abstract: Taking a DNA sequence, a word with letters/bases A, T, G and C, as the relation between the generators of an infinite group \(\pi\), one can discriminate between two important families: (i) the cardinality structure for conjugacy classes of subgroups of \(\pi\) is that of a free group on one to four bases, and the DNA word, viewed as a substitution sequence, is aperiodic; (ii) the cardinality structure for conjugacy classes of subgroups of \(\pi\) is not that of a free group, the sequence is generally not aperiodic and topological properties of \(\pi\) have to be determined differently. The two cases rely on DNA conformations such as A-DNA, B-DNA, Z-DNA, G-quadruplexes, etc. We found a few salient results: Z-DNA, when involved in transcription, replication and regulation in a healthy situation, implies (i). The sequence of telomeric repeats comprising three distinct bases most of the time satisfies (i). For two-base sequences in the free case (i) or non-free case (ii), the topology of \(\pi\) may be found in terms of the \(SL(2, \mathbb{C})\) character variety of \(\pi\) and the attached algebraic surfaces. The linking of two unknotted curves—the Hopf link—may occur in the topology of \(\pi\) in cases of biological importance, in telomeres, G-quadruplexes, hairpins and junctions, a feature that we already found in the context of models of topological quantum computing. For three- and four-base sequences, other knotting configurations are noticed and a building block of the topology is the four-punctured sphere. Our methods have the potential to discriminate between potential diseases associated to the sequences.

Keywords: DNA conformations; transcription factors; telomeres; infinite groups; free groups; algebraic surfaces; aperiodicity; character varieties

1. Introduction

Group theory and algebraic geometry serve the decipherment of ‘the book of life’ [1], a book made of a language employing four letters/nucleotides: A (adenine), T (thymine), G (guanine) and C (cytosine), as described in this work. There are finite groups, groups made of a finite number of generators and a finite number of elements that may be used to map the codons to amino acids, as carried out in our papers [2,3]. Such an approach toward the genetic code is made possible by identifying the irreducible characters of the group to the amino acids. The multiplets of codons attached to a selected amino acid correspond to the irreducible characters having the corresponding dimension of the representation (Table 3 in [2], Table 4 in [3]). A virtue of the approach is that the used irreducible characters are also seen as quantum states carrying complete quantum information.

For modeling DNA in its various conformations taken in transcription factors, telomeres and other building blocks of molecular biology, we need infinite groups defined from a motif. A sequence of the DNA nucleotides serves as the generator of the group [4]. In this context, it has been found that a group that is not free is often the witness of a potential disease. We coined the term ‘syntactical freedom’ for recognizing this property, with inspiration from an earlier work [5]. We also showed that such free groups have the distinctive
property of generating an aperiodic substitution rule, providing a connection between (group) syntactical freedom and irrational numbers (Section 4 in [4]). For an infinite group, the representation cannot be based on characters but on the so-called character variety. This topic leads to a relationship between DNA, algebraic topology and algebraic geometry. Tools already proposed for topological quantum computing [6] are also used in the context of DNA conformations.

In Section 2, we briefly account for the many types of topologies that DNA can show, in terms of double strands or more strands. Then, we recall the mathematical concepts employed in our paper with some redundancy with earlier work [4,6].

In Section 3, we explain the concept of an $SL(2,\mathbb{C})$ character variety associated to an infinite group with two or three generators. The former case corresponds to DNA motifs having only two distinct nucleotides. In such a case, the variety often contains the Cayley cubic associated to the Hopf link, the non-disjoint union of two circles in the three-dimensional space. In the latter case, the variety contains the Fricke–Klein seventh variable polynomial that is characteristic of the topology of the three-dimensional sphere with four points removed.

In Section 4, we apply these mathematical methods to transcription factors, telomeric sequences and a specific DNA decamer sequence, where almost all of its conformations have been crystallized.

2. Materials and Methods

Mathematical calculations performed in this paper are on the software Magma [7] (for groups) or on Sage software [8] (for character varieties).

2.1. DNA Conformations

DNA is a long polymer made from a chain of the nucleotides A, T, C or G. DNA exists in many possible conformations, which include a double-stranded helix of A-DNA, B-DNA and Z-DNA, although only B-DNA and Z-DNA have been directly observed in functional organisms [9,10]. The B-DNA form is most common under the conditions found in cells, but Z-DNA is often preferred when DNA binds to a protein. A view of a double helix in the A-, B- and Z-DNA forms is given in Figure 1. Other DNA conformations also exist, such as a single-stranded hairpin used mostly in macromolecular synthesis and repair, a triple-stranded H-DNA found in peptides, a G-quadruplex structure found in telomeres and a Holliday junction.

![Figure 1](image-url). From left to right, the structures of A-, B- and Z-DNA. The view of the double helix from above (or below) shows distinct symmetries, 11-fold for the A-DNA, 10-fold for the B-DNA and 6-fold for the Z-DNA [9,10].
2.2. Finitely Generated Groups, Free Groups and Their Conjugacy Classes, and Aperiodicity of Sequences

The free group $F_r$ on $r$ generators (of rank $r$) consists of all distinct words that can be built from $r$ letters where two words are different unless their equality follows from the group axioms. The number of conjugacy classes of $F_r$ of a given index $d$ is known and is a good signature of the isomorphism, or the closeness, of a group $\pi$ to $F_r$. In the following, the cardinality structure of conjugacy classes of index $d$ in $F_r$ is called the cardinality sequence (card seq) of $F_r$, and we need the cases from $r = 1$ to $3$ to correspond to the number of distinct bases in a DNA sequence. The card seq of $F_r$ is in Table 1 for the three sequences of interest in the context of DNA [11].

Table 1. Number of conjugacy classes of subgroups of index $d$ in free group of rank $r = 1$ to $3$ [11]. The last column is the index of the sequence in the on-line encyclopedia of integer sequences [12].

| $r$ | Card Seq | Sequence Code |
|-----|----------|---------------|
| 1   | [1, 1, 1, 1, 1, 1] | A000012 |
| 2   | [1, 3, 7, 26, 97, 624, 4163, 34470, 314493] | A057005 |
| 3   | [1, 7, 41, 604, 13753, 504243, 24824785, 1598346352] | A057006 |

Next, given a finitely generated group $f_p$ with a relation (rel) given by the sequence motif, we are interested in the card seq of its conjugacy classes. Often, the DNA motif in the sequence under investigation is close to that of a free group $F_r$, with $r + 1$ being the number of distinct bases involved in the motif. However, the finitely generated group $f_p = \langle x_1, x_2 | rel(x_1, x_2) \rangle$ or $f_p = \langle x_1, x_2, x_3 | rel(x_1, x_2, x_3) \rangle$ or $f_p = \langle x_1, x_2, x_3, x_4 | rel(x_1, x_2, x_3, x_4) \rangle$ (where the $x_i$ are taken in the four bases A, T, G and C, and rel is the motif), may not be the free group $F_1 = \langle x_1, x_2 | x_1x_2 \rangle$, or $F_2 = \langle x_1, x_2, x_3, x_4 | x_1x_2x_3 \rangle$ or $F_3 = \langle x_1, x_2, x_3, x_4 | x_1x_2x_3x_4 \rangle$. The closeness of $f_p$ to $F_r$ can be checked by its signature in the finite range of indices of the card seq.

2.2.1. Groups $f_p$ Close to Free Groups and Aperiodicity of Sequences

According to reference [5], aperiodicity correlates to the syntactical freedom of ordering rules. This statement was checked in the realm of transcription factors (Section 4 in [4]). Let us introduce the concept of a general substitution rule in the context of free groups. A general substitution rule $\rho$ on a finite alphabet $A_r$ on $r$ letters is an endomorphism of the corresponding free group $F_r$ (Definition 4.1 in [13]). The endomorphism property means the two relations $\rho (uv) = \rho (u) \rho (v)$ and $\rho (u^{-1}) = (\rho (u))^{-1}$, for any $u, v \in F_r$.

A special role is played by the subgroup $\text{Aut}(F_r)$ of automorphisms of $F_r$. We introduce the map $\alpha : F_r \to \mathbb{Z}_r$ from $F_r$ to the Abelian group $\mathbb{Z}_r$ in order to investigate the substitution rule $\rho$ with the tools of matrix algebra.

The map $\alpha$ induces a homomorphism $M : \text{End}(F_r) \to \text{Mat}(r, \mathbb{Z})$. Under $M$, $\text{Aut}(F_r)$ maps to the general linear group of matrices with integer entries $GL(r, \mathbb{Z})$. Given $\rho$, there is a unique mapping $M(\rho)$ that makes the map diagram commutative [13] (p. 68). The substitution matrix $M(\rho)$ of $\rho$ may be specified by its elements at row $i$ and column $j$ as follows:

$$ (M(\rho))_{ij} = \text{card}((\rho_a(a_j))). $$

This approach was applied to binding motifs of transcription factors [4]. The binding motif $\text{rel}$ in the finitely presented group $f_p = \langle A, T, G, C | \text{rel}(A, T, G, C) \rangle$ is split into appropriate segments so that $\text{rel} = \text{rel}_A \text{rel}_T \text{rel}_G \text{rel}_C$ with the substitution rules $A \to \text{rel}_A$, $T \to \text{rel}_T$, $G \to \text{rel}_G$, $C \to \text{rel}_C$.

We are interested in the sequence of finitely generated groups

$$ f_p^{(l)} = \langle A, T, G, C | \text{rel}(\text{rel} \cdots (A, T, G, C) \rangle \rangle \quad \text{(with rel applied l times)} $$


whose card seq is the same at each step \( l \) and equal to the card seq of the free group \( F_l \) (in the finite range of indices that it is possible to check with the computer).

Under these conditions, (group) syntactical freedom correlates to the aperiodicity of sequences.

### 2.2.2. Aperiodicity of Substitutions

There is no definitive classification of aperiodic order, the intermediate between crystalline order and strong disorder, but in the context of substitution rules, some criteria can be found. First, we need a few definitions.

A non-negative matrix \( M \in \text{Mat}(d, \mathbb{R}) \) is one whose entries are non-negative numbers. A positive matrix \( M \) (denoted \( M > 0 \)) has at least one positive entry. A strictly positive matrix (denoted \( M >> 0 \)) has all positive entries. An irreducible matrix \( M = (M_{ij})_{1\leq i,j\leq d} \) is one for which there exists a non-negative integer \( k \) with \( (M^k)_{ij} > 0 \) for each pair \( (i,j) \).

A primitive matrix \( M \) is one such that \( M^k \) is a strictly positive matrix for some \( k \).

A Perron–Frobenius (PF for short) eigenvector \( v \) of an irreducible non-negative matrix is the only one whose entries are positive: \( v > 0 \). The corresponding eigenvalue is called the PF eigenvalue.

We will use the following criterion (Corollary 4.3 in [13]). A primitive substitution rule \( \rho \) of substitution matrix \( M(\rho) \) with an irrational PF-eigenvalue is aperiodic.

A well-established primitive substitution rule is the Fibonacci rule \( \rho_F \) where \( \rho_{FG} : a \rightarrow ab, b \rightarrow a \) of substitution matrix \( M_F = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \) and PF-eigenvalue equal to the golden ratio \( \lambda_{PF} = \varphi = (\sqrt{5} + 1)/2 \) (Example 4.6 in [13]). As expected, the irrationality of \( \lambda \) corresponds to the aperiodicity of the Fibonacci sequence.

The sequence of Fibonacci words is as follows:

\[ a, b, ab, aba, abaab, abaabaabaab, \cdots \]

The words have lengths equal to the Fibonacci numbers \( 1, 1, 2, 3, 5, 8, 13, 21, \cdots \)

All finitely generated groups \( f_p^{(l)} \) whose relations \( \text{rel}(a,b) = ab, aba, abaab, \cdots \) have a card seq whose elements are 1s, as for the card seq of the free group \( F_1 \). The Fibonacci sequence is our first example where group syntactical freedom correlates to aperiodicity.

### 2.2.3. A Four-Letter Sequence for the Transcription Factor of the Fos Gene

Let us now apply the method to a transcription factor of importance. The transcription factor of gene Fos has selected motif \( \text{rel} = T G A G T C A \) [14]. For this case, the four-letter generated group has a card seq similar to the free group \( F_3 \) given in Table 1.

We split \( \text{rel} \) into four segments so that \( \text{rel} = \text{rel}_A \text{rel}_T \text{rel}_G \text{rel}_C \) with the substitution maps \( A \rightarrow \text{rel}_A = T, T \rightarrow \text{rel}_T = G, G \rightarrow \text{rel}_G = ATGC, C \rightarrow \text{rel}_C = A \) to produce the substitution sequence

\[ A, T, G, C, ATGC, TGAGTCA, GAGTCTAGTCGAT \cdots \]

The substitution matrix for this sequence is \( M = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \). It is a primitive matrix \( (M^4 >> 0) \) whose eigenvalues follow from the vanishing of the polynomial \( \lambda^4 - \lambda^3 - \lambda^2 - \lambda - 1 \). There are two real eigenvalues \( \lambda_1 \approx 1.92756 \) and \( \lambda_2 \approx -0.77480 \), as well as two complex conjugate eigenvalues \( \lambda_{3,4} \approx -0.07637 \pm 0.81470i \). The PF-eigenvalue is \( \lambda_{PF} = \lambda_1 \), with an eigenvector of (positive) entries \( \approx (1, 0.37298, 0.40211, 0.20861)^T \). It follows that the selected sequence for the Fos gene is aperiodic.

All of the finitely generated groups \( f_p^{(l)} \) whose relations are

\[ \text{rel}(A, C, G, T) = ATGC, TGAGTCA, GAGTCTAGTCGAT, \cdots \]
have a card seq whose elements are

\[1, 7, 41, 604, 13753, 504243, \cdots,\]

which is the card seq of the free group \(F_3\). For the Fos transcription factor, group syntactical freedom correlates to aperiodicity as expected.

Further examples are obtained in the context of DNA sequences for transcription factors (Section 4 in [4]) and below, in relation to DNA conformations and telomeres.

3. Discussion

In the following, we make use of \(SL(2, \mathbb{C})\) representations of the infinite groups \(\pi\) arising from specific DNA sequences. The character variety has many interpretations in mathematics and physics. For instance, in mathematics, the variety is the space of representations of hyperbolic structures of three-manifolds \(M\) with fundamental group \(\pi(M)\), and the variety of the characters of \(SL(2, \mathbb{C})\) representations of \(\pi(M)\) is reflected in the algebraic geometry of the character variety [15,16]. In physics, the group \(SL(2, \mathbb{C})\) expresses the symmetries of fundamental physical laws. It is also known as the Lorentz group; more precisely, the double cover of the restricted Lorenz group is \(SL(2, \mathbb{C})\), which is the spin group.

3.1. \(SL(2, \mathbb{C})\) Character Varieties and Algebraic Surfaces

Recently, we found that the representation theory of finite groups with their character table allows us to derive an approach of the genetic code [3].

For infinite groups \(\pi\) such as those defined by DNA sequences, it is useful to describe the representations of \(\pi\) in the Lorentz group \(SL(2, \mathbb{C})\), the group of \((2 \times 2)\) matrices with complex entries and determinant 1. Such a group expresses the fundamental symmetry of all known physical laws, apart from gravitation.

Representations of \(\pi\) in \(SL(2, \mathbb{C})\) are homomorphisms \(\rho: \pi \to SL(2, \mathbb{C})\) with character \(\kappa_{\rho}(g) = \text{tr}(\rho(g)), g \in \pi\). The set of characters allows us to define an algebraic set by taking the quotient of the set of representations \(\rho\) by the group \(SL_2(\mathbb{C})\), which acts by conjugation on representations [15,17].

For two-generator groups, the character variety may be decomposed into the product of surfaces, which reveals the topology of \(M\). We recently found a connection between some groups whose topology is based on the Hopf link and a model of topological quantum computing [6]. The Hopf link underlies many DNA sequences whose group structure is (or is not) that of the free group \(F_1\). The classification of the involved algebraic surfaces in the variety is performed using specific tools available in Magma [7]; see (Section 2.1 in [6]) for details.

For three-generator groups, we find that the Fricke–Klein quartic is part of the character variety.

3.2. The Hopf Link

Taking the linking of two unknotted curves as in Figure 2 (Left), the obtained link is called the Hopf link \(H = L2a1\), whose knot group is defined as the fundamental group of the knot complement in the three-sphere \(S^3\)

\[\Pi_1(S^3 \setminus L2a1) = \langle a, b | [a, b] \rangle = \mathbb{Z}^2,\]  

where \([a, b] = abAB\) (with \(A = a^{-1}\), \(B = b^{-1}\)) is the group theoretical commutator.
Figure 2. (Left): the Hopf link. (Right): the link $L = A \cup B$ is attached to the plane $R^2$ in the half-space $R^3_+$. It is not splittable. This can be proved by checking that the fundamental group $\pi = \pi_2(L)$ is not free [18] and p. 90 in [19]. One gets $\pi_2 = \langle x,y,z|(x,y,z) = z \rangle$, where $(..)$ means the group theoretical commutator. The cardinality sequence of $cc$ of subgroups of $\pi_2$ is $[1,3,10,51,164,1365,9422,81594,721305,\cdots]$ (Figure 3 in [4]).

There are interesting properties of the knot group $\Pi_1$ of the Hopf link. First, the number of coverings of degree $d$ of $\Pi_1$ (which is also the number of conjugacy classes of index $d$) is precisely the sum of divisor function $\sigma(d)$ [20].

Second, an invariance of $\Pi_1$ under a repetitive action of the golden ratio substitution (the Fibonacci map) $\rho : a \to ab$, $b \to a$ or under the silver ratio substitution $\rho : a \to aba$, $b \to a$ exists. The terms golden and silver refer to the Perron–Frobenius eigenvalue of the substitution matrix (Examples 4.5 and 4.6 in [13]). Such an observation links the Hopf link, the group $\Pi_1$ of the 2-torus and aperiodic substitutions.

Using Sage software [8] developed from Ref. [17], the $SL_2(C)$ character variety is the polynomial corresponding to the so-called Cayley cubic

$$f_H(x,y,z) = xyz - x^2 - y^2 - z^2 + 4. \quad (2)$$

As expected, the three-dimensional surface $\Sigma : f_H(x,y,z) = 0$ is the trace of the commutator and is known to correspond to the reducible representations (Theorem 3.4.1 in [21]). A picture is given in Figure 3 (left).

Figure 3. (Left): a three-dimensional picture of the $SL_2(C)$ character variety $\Sigma_H$ for the Hopf link complement $H$. (Right): a modified character variety of defining equation $f_\tilde{H}(x,y,z)$ with similar singularities.

In the perspective of algebraic geometry, we classify the homogenization of equation $f_H$ as a rational surface of degree 3 del Pezzo type. It displays four simple singularities.
3.3. Beyond the Hopf Link

As shown in [6], the Hopf link is the irreducible component of many character varieties relevant to a model of topological quantum computing. In the context of DNA groups investigated in the next section, we also find another surface with similar simple singularities as shown in Figure 3 (right). The defining polynomial is

\[ f_H(x, y, z) = z^4 - 2xyz + 2x^2 + 2y^2 - 3z^2(-4z) - 4. \]  

(3)

The homogenization of equation \( f_H(x, y, z) \) allows us to classify it as a conic bundle in the family of \( K_3 \) surfaces.

For the DNA sequence, whose group \( \pi_H \) contains the component \( f_H(x, y, z) \), we refer to the third subsection of the Results section below and the first table in this subsection. The relevant triplet \( nnn=CGG \) of the dodecameric sequence \( d(CCCCCGCGGGGG) \) provided in [4], there is the additional fourth column that signals when the Groebner base to the free group

\[ \langle \alpha, \beta, \gamma \rangle \] \[ \alpha \beta \gamma \delta \]

The Cayley cubic is a subset of the character variety for the four-punctured three-dimensional sphere \( S^2_4 \) (the sphere minus four points). Its fundamental group \( \Pi_1 \) is isomorphic to the free group \( F_3 \) of rank 3, \( \Pi_1(S^2_4) = \langle a, b, \gamma, \delta \rangle \); \( \alpha \beta \gamma \delta \) corresponds to loops around the punctures.

The \( SL(2, \mathbb{C}) \) character variety for \( \Pi_1(S^2_4) \) satisfies a quartic equation in terms of the Fricke–Klein seventh variable polynomial [21] (p. 65) and [22]:

\[ f(x, \theta) = xyz + x^2 + y^2 + z^2 - \theta_1 x - \theta_2 x - \theta_3 z + \theta_4, \]

(4)

with \( \theta_1 = uv + wk, \theta_2 = uw + vk, \theta_3 = uk + vw, \theta_4 = uvwk + u^2 + v^2 + w^2 - 4. \)

4. Results

In this section, we apply the \( SL(2, \mathbb{C}) \) representation theory to specific non-canonical DNA sequences having regulatory functions in gene expression (the transcription factors), replication (the telomeres) and DNA conformations.

4.1. Group Structure and Topology of Transcription Factors

In a transcription factor, a motif-specific DNA binding factor controls the rate of the transcription of a gene from DNA to messenger RNA by binding a protein to the DNA motif. In reference [4], we found a correlation between motifs whose subgroup structure is that of a free group and the lack of a potential disease while the gene is activated in transcription, the property of ‘syntactical freedom’.

In Table 2, this idea is illustrated by restricting to a few transcription factors whose motif comprises two bases. The card seq of the motif is either the free group \( F_1 \), close to \( F_3 \) or away from a free group when the card seq is that of the modular group \( H_3 \), of the Baumslag–Solitar group \( BS(-1, 1) \) or that of groups \( \pi_4 \) and \( \pi_4' \). Compared to the results provided in [4], there is the additional fourth column that signals when the Groebner base for the ideal ring of the \( SL(2, \mathbb{C}) \) character variety contains the Cayley cubic, the unique component in the case of the Hopf link [6], a degree 3 del Pezzo surface (denoted HL), or not. An additional fifth column is filled to check the presence of a surface of type \( K_3 \). Only the last row of the table for the transcription factor of gene EHF does not show this property.
Table 2. Group structure of motifs for a few two-letter transcription factors. The card seq for the modular group $H_3$ is $[1, 1, 2, 3, 2, 8, 7, 10, 18, 28, \cdots]$. The Baumslag–Solitar group $BS(-1, 1)$ is the fundamental group of the Klein bottle. The card seq for $BS(-1, 1)$ is $[1, 3, 2, 5, 2, 7, 2, 8, 1, 2, 7, 2, 3, \cdots]$; for $\pi_1'$, it is $[1, 1, 1, 2, 1, 3, 2, 1, 1, 9, 2, 14, 2, 1, \cdots]$. The symbol HL means that the Cayley cubic is part of the Groebner base for the ideal ring of the corresponding $SL_2(C)$ character variety.

For three-letter transcription factors, the ideal ring of the corresponding $SL_2(C)$ character variety contains the Fricke–Klein seventh variable polynomial 4, which is a feature of the four-punctured sphere topology. The group structure of three-letter transcription factors not leading to free groups is shown in (Table 5 in [4]).

| Gene     | Motif     | Card Seq | Link | Type | Literature       |
|----------|-----------|----------|------|------|------------------|
| DBX      | TTTATTA   | $F_1$    | HL   | $K_3$| [23], MA0174.1  |
| SPT15    | TATATATT  | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| PHOX2A   | TAATTAAATA| $\approx F_1$ | $\cdots$ | $\cdots$ | $\cdots$ |
| FOXA     | TGTGTGTGT | $F_1$    | $\cdots$ | $\cdots$ | $\cdots$ |
| FOXG     | TTTGTGTG  | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| NKX6-2   | TAATTAA   | $H_3$    | no   | $K_3$| [23], [MA0675.1, MA0675.2] |
| FOXG     | TGTGTG    | $BS(-1, 1)$ | no   | $K_3$| [23,26], MA1865.1 |
| HoxA1, HoxA2 | TAATTA | $\pi_1$ | no   | $K_3$| [23], [MA1495.1, MA0900.1] |
| POU6f1, Vax | TGTGTT | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| RUNX1    | TGTGTT    | $\pi_1'$ | no   | $K_3$| [23], [MA0628.1, MA0722.1] |
| RUNX1    | TGTGTT    | $\pi_1'$ | no   | $K_3$| [23, MA0002.2] |
| EHF      | CCTCCCTC  | $F_1$    | HL   | $K_3$| $\cdots$, MA0598.1 |

The Character Variety for the Transcription Factor of the DBX Gene

We explicitly show the $SL_2(C)$ character variety for the transcription factor of the DBX gene.

$$f_{DBX}(x,y,z) = f_H(x,y,z)(yz^2 - y^2 - xz - y + 2)(xy^2 - z^2 - yz - x + z)$$

$$= (y^3 - z^2 - 3y + 2)(y^2z - xy - yz + x - z)(z^4 - x^2y + xz - 4z^2 + y + 2)$$

(5)

The factors in (5) are three degree 3 del Pezzo surfaces (including the Cayley cubic $f_H$), two rational ruled surfaces and a $K_3$ surface birationally equivalent to the projective plane, respectively. The latter factor also belongs to the character variety of group $\Pi_1(S^4 \setminus \tilde{E}_6)$, where $S_4$ is the four-sphere and $\tilde{E}_6$ is the singular fiber $IV^*$ in Kodaira’s classification of minimal elliptic surfaces (Figure 4b in [6]).

It is important to mention that, for three-letter transcription factors, the ideal ring of the corresponding $SL_2(C)$ character variety contains the Fricke–Klein seventh variable polynomial (4), which is a feature of the four-punctured sphere topology.

Table 3 provides a short account of the function or potential dysfunction of the genes under consideration. As mentioned before, most of the time, such a dysfunction is correlated to a card seq away from that of the free group $F_1$.

In view of our results, it is interesting to correlate the presence of the Hopf link HL in the character variety with the possible remodeling of B-DNA into Z-DNA or another DNA conformation. To our knowledge, general information about this subject is still lacking. From a biological point of view, it is known that some of the Z-DNA-forming conditions that are relevant in vivo are the presence of DNA supercoiling, Z-DNA-binding proteins [27] and base modifications. When transcription occurs, the movement of RNA polymerase II along the DNA strand generates positive supercoiling in front of, and negative supercoiling behind, the polymerase [28].
Table 3. A short account of the function or dysfunction (through mutations or isoforms) of genes associated with transcription factors and sections in Table 2.

| Gene     | Type       | Function                                      | Dysfunction                      |
|----------|------------|-----------------------------------------------|----------------------------------|
| DBX      | TATA-box   | drosophila segmentation                       |                                  |
| SPT15    | binding protein | gene expression, regulation                | fibrosis of extraocular muscles |
| PHOX2A   | homeodomain | differentiation, maintenance of noradrenergic phenotype | diabetes                        |
| FOX proteins | forkhead box | growth, differentiation, insulin secretion longevity | spastic ataxia                  |
| FOXA2    |            |                                               |                                  |
| NKX6-2   | homeobox   | central nervous system, pancreas             | spastic ataxia                  |
| FOXG     | forkhead box | notochord (neural tube)                    | chordoma                        |
| HoxA1    | homeobox   | embryonic devt of face and hear             | autism                          |
| HoxA2    |            | embryonic devt of face and hear             | cleft palate                    |
| Pou6F1   | .          | neuroendocrine system                        | clear cell adenocarcinoma        |
| Vax      | .          | forebrain development                       | craniofacial malform.           |
| RunX1    | Runt-related | cell differentiation, pain neurons        | myeloid leukemia                 |
| EHF      | homeobox   | epithelial expression                      | carcinogenesis, asthma          |

Perhaps the lack of HL in the character variety for transcription factors of genes in Table 2 means that the Z-DNA-forming condition is not realized.

4.2. Group Structure and Topology of DNA Telomeric Sequences

Terminal structures of chromosomes are made of short highly repetitive G-rich sequences with proteins known as telomeres. They have a protective role against the shortening of chromosomes through successive divisions. Most organisms use a telomere-specific DNA polymerase called telomerase that extends the 3’ end of the G-rich strand of the telomere [29]. Telomere shortening is associated with aging, mortality and aging-related diseases such as cancer.

A list of results obtained by using our group theoretical approach is in Table 4. For two-letter telomere sequences, the SL(2, C) character variety contains the Cayley cubic, the characteristic of the Hopf link HL, only in the first row. In addition to the Cayley cubic, one finds surfaces of a general type. In the next two rows, the Cayley cubic is not found. There are degree 3 del Pezzo surfaces in the factors of the character variety but not general surfaces.

As for the Hopf link, the sequence is found to be aperiodic with the Perron–Frobenius eigenvalue $\lambda_{PF}$ equal to the golden ratio. For three-letter telomere sequences, the card seq is that of the free group of $F_2$, except for the last row, where the identified group is $\pi_2$; see Figure 2 (right) for the definition of such a group. In the former seven cases, the DNA topology is known to be a G-quadruplex structure [30–36]. We could identify an aperiodic structure of the telomere sequence with the Perron–Frobenius eigenvalue $\lambda_{PF}$ as shown in column 5. In the latter case, the topology is of the basket type [36] and no aperiodicity of the telomere sequence could be found.

Figure 4, taken from the protein data bank (PDB 2HY9), illustrates the G-quadruplex structure of the telomere sequence in vertebrates.
Table 4. Group analysis of the telomere sequence found in some eukaryotes. The first column is for the telomere repeat, the second column is the organism under investigation, the third column is for the PDB code, the fourth column is for the card seq of the group $\pi$ or that of the corresponding group that is identified, the fifth column is for the Perron–Frobenius eigenvalue when the sequence is found to be aperiodic, the sixth column identifies the presence of the Hopf link (in two-base sequences) or the DNA conformation (in three-base sequences) and the seventh column is a relevant reference. The notation G-quadr is for the G-quadruplex; see Figure 4. The card seq for $\pi_1''$ is $[1, 3, 2, 16, 16, 69, 118, 719, 1877, 8949 \cdots]$. The Hecke group $H_4$ is defined in (Table 2 in [4]).

| Seq       | Organism       | PDB   | Card Seq | $\lambda_{PF}$ | Link/DNA Conf | Ref   |
|-----------|----------------|-------|----------|----------------|---------------|-------|
| G4T4G4    | Oxytricha      | 1D59  | $\pi_1''$ | $(\sqrt{5} + 1)/2$ | HL            | [37]  |
| TG4T      | universal      | 244D.1| $H_4$    | .              | no            | [38]  |
| TG4       | Tetrahymena    | 230D  | $H_4$    | .              | no            | [29]  |
| T2AG3     | Vertebrates    | 2HY9  | $F_2$    | 2.5468         | G-quadr.      | [30]  |
| TAG3      | Giardia        | 2KOW  | $F_2$    | 2.2055         | G-quadr        | [31]  |
| T2AG2     | Bombyx mori    | unknown| $F_2$    | no             | G-quadr        | [32]  |
| T4AG3     | Green algae    | unknown| $F_2$    | 3.07959        | unknown        | [33]  |
| G2TAG2    | Human          | 2HRI  | $F_2$    | 2.5468         | G-quadr        | [34]  |
| TAGT2AG3  | Human          | 2HRI  | $F_2$    | 3.3923         | G-quadr        | [35]  |
| G3TAG3    | Human          | unknown| $F_2$    | 4.3186         | G-quadr        | [36]  |
| G3TAG3    | Human          | unknown| $\pi_2$  | no             | basket         | [36]  |
| (GGGTTA)3G3T| Human      | unknown| unknown  | .              | no            | [36]  |

Figure 4. Human telomere DNA quadruplex structure in K+ solution hybrid-1 form, PDB 2HY9 [30].

4.3. Group Structure and Topology of the DNA Decamer Sequence $d(CCnnnN_nN_nN_nGG)$ [10]

A challenging question of structural biology is to determine if and how a DNA (or RNA) sequence defines the three-dimensional conformation, as well as the secondary and tertiary structure of proteins. In the previous two subsections, we tackled the problem with regard to transcription factors and telomeric sequences, respectively. In the former case, we restricted to the DNA part of the transcription since the DNA motif is almost exactly known from X-ray techniques while the secondary structure of the binding protein strongly depends on the model employed and the choice made to recognize the sections of the secondary structures (e.g., alpha helices, beta sheets and coils) [39]. In the latter case, in many organisms, nature invented telomerase for taking care of the replication without damaging the sequences at the 3-ends too much, while keeping the catalyzing action of DNA polymerase. Again, there is a loop complex in telomerase comprising telomere-binding proteins, with secondary structures not being analyzed so far with our group theoretical approach.
In this section, we also study DNA conformations and their relationship to algebraic topology in a specific DNA decamer sequence investigated in reference [10] by a standard crystallization technique followed by X-ray diffraction discrimination. In the sequence \( d(C{\text{mnnN}}_6{\text{nN}}_7{\text{nG}}_8) \), the factors \( N_6, N_7 \) and \( N_8 \) are taken in the two nucleotides \( G \) and \( C \), and \( nnn \) is specified in order to maintain the self-complementarity of the sequence. This inverse repeated motif is the minimum motif used to distinguish between the double-strand forms of \( B \)- and \( A \)-DNA, while excluding the \( Z \)-DNA forms. A third conformation is allowed and called the four-stranded Holliday junction \( J \). We refer to (Table 1 in [10]) for the main results.

On our side, the card seq of each sequence was determined and the \( SL(2, \mathbb{C}) \) character variety was obtained. Our results are summarized in the four Tables 5–8.

**Table 5.** Group analysis of the sequence \( d(C{\text{mnnN}}_6{\text{nN}}_7{\text{nG}}_8) \), where \( N_6, N_7 \) and \( N_8 \) are taken in the two nucleotides \( G \) and \( C \) and \( nnn \) is specified in order to maintain the self-complementarity of the sequence [10]. The first column is for the selected triplet \( N_6N_7N_8 \), the second column is for the code in the protein data bank, the third column is for the DNA conformation when known (see Table 1 in [10]), the fourth column is for the cardinality structure of subgroups of \( \pi \) and the fifth column checks the occurrence of a surface corresponding to the Hopf link in the factorization of the \( SL(2, \mathbb{C}) \) of \( \pi \). The symbols \( A, B \) and \( J \) are for \( A \)-DNA, \( B \)-DNA and a four-stranded Holliday junction; lowercase is used when the conformation is not confirmed in [10].

| Triplet | PDB | Conformation | Card Seq (\( \pi \)) | Knot |
|---------|-----|--------------|----------------------|------|
| CCC     | 1ZF1 | A            | [1,1,1,7,1,1,2,9,6,\ldots] | HL   |
| CCC     | 1ZF2 | J            | idem                 | HL   |
| CCG     | 1ZEX | A            | idem                 | HL   |
| CCG     | 1ZEY | A            | [1,1,2,6,3,1,4,2,6,\ldots] | HL like |
| CCG     | none | unknown      | [1,1,2,1,6,3,2,1,3,6,\ldots] | no   |
| GGG     | 1ZF9 | A            | [1,1,1,10,25,25,9,2,1798,\ldots] | no   |
| GCC     | none | b/J          | [1,1,1,1,6,1,2,1,1,6,\ldots] | HL   |
| GCC     | none | unknown      | [1,1,2,2,7,5,1,4,5,9,\ldots] | no   |
| GCC     | none | B/a          | [1,1,1,1,6,11,9,5,2,208,\ldots] | no   |

(card seq of Hecke group \( H_5 \))

**Table 6.** Group analysis of the sequence \( d(C{\text{mnnN}}_6{\text{nN}}_7{\text{nG}}_8) \), where \( N_6, N_7 \) and \( N_8 \) are taken in the two nucleotides \( A, T \) [10]. Groups \( \pi_3 \) and \( \pi_4 \) are as in (Table 5 in [4]). The card seq for \( \pi_3' \) is \( [1,7,50,867,15906,570528,\ldots] \); for \( \pi_4' \), it is \( [1,7,50,739,15234,548439,\ldots] \); for \( \pi_4'' \), it is \( [1,7,59,1258,24787,\ldots] \). Groups \( \pi_3' \) and \( \pi_4'' \) may be simplified to a group whose card seq is that of \( \pi_2 \), the fundamental group of the link \( L = A \cup B \) described in Figure 3 (right).

| Triplet | PDB | Conformation | \( \pi \) |
|---------|-----|--------------|---------|
| TTA     | 1ZFH | B            | \( \pi_3' \rightarrow \pi_2 \) |
| TAA     | none | B            | \( \pi_3' \rightarrow \pi_2 \) |
| ATT     | none | b            | \( \pi_3' \rightarrow \pi_2 \) |
| AAA     | none | unknown      | \( \pi_3' \rightarrow \pi_2 \) |
| TTT     | none | unknown      | \( \pi_3' \rightarrow \pi_2 \) |
| ATA     | none | unknown      | \( \pi_3' \rightarrow \pi_2 \) |
| TAT     | none | unknown      | \( \pi_3' \rightarrow \pi_2 \) |

In Table 5, \( N_6, N_7 \) and \( N_8 \) are taken in the two nucleotides \( G \) and \( C \), forming eight triplets and the associated two-letter decamer sequences. Note that the triplet CCC produces two distinct DNA conformations \( A \) and \( J \). The character variety of the Hopf link HL (the Cayley cubic) is present in the factors of the ideal ring of the \( SL(2, \mathbb{C}) \) character variety in five cases over the nine possibilities, where one case (with triplet CGG and code 1ZEY in the protein data bank) shows an algebraic surface similar to the Cayley cubic (with four
simple singularities) as defined in Equation (3) and as shown in Figure 3 (right). We do not observe a clear correlation between the type of DNA conformation and the underlying HL topology, but the presence of HL in the variety seems to exclude the B-DNA conformation. In addition, the character variety always contains a surface of type $K_3$ in its factors.

Table 7. Group analysis of the sequence $d(\text{CCmnnN}_6 \text{N}_7 \text{N}_8 \text{GG})$ [10], where $\text{N}_6$, $\text{N}_7$ and $\text{N}_8$ are taken in the two nucleotides $\text{A}, \text{G}$ (left) and $\text{A}, \text{C}$ (right). Groups $\pi_3$ and $\pi_4$ are as in (Table 5 in [4]). The card seq for $\pi_3^{(3)}$ is $[1, 7, 61, 169, 668, 14969, \cdots]$ and, for $\pi_3^{(5)}$, it is $[1, 7, 41, 604, 28153, \cdots]$.

| Triplet | PDB | Conformation | $\pi$ | Triplet | PDB | Conformation | $\pi$ |
|---------|-----|--------------|-------|---------|-----|--------------|-------|
| AGA     | 1ZEW| B            | $F_3$ | ACA     | none | unknown      | $\pi_3^{(3)} \rightarrow \pi_2$ |
| AGG     | none | unknown      | $\pi_3^{(5)}$ | ACC     | none | J           | $\pi_3$ |
| GGA     | 1ZFA| A            | $F_3$ | CCA     | none | unknown      | $\pi_3$ |
| AAG     | none | unknown      | $F_3$ | AAC     | 1ZF0 | B           | $\pi_3^{(5)} \rightarrow \pi_2$ |
| TGT     | none | unknown      | $F_3$ | TCT     | none | b           | $\pi_3^{(3)} \rightarrow \pi_2$ |
| TGG     | 1ZF6| A            | $F_3$ | TCC     | none | unknown      | $\pi_3$ |
| GGT     | 1ZF8| A            | $F_3$ | CCT     | none | b           | $\pi_3$ |
| TTG     | none | unknown      | $\pi_3^{(5)}$ | TTC     | none | b           | $\pi_3^{(5)} \rightarrow \pi_2$ |

Table 8. Group analysis of the sequence $d(\text{CCmnnN}_6 \text{N}_7 \text{N}_8 \text{GG})$ [10], where $\text{N}_6$, $\text{N}_7$ and $\text{N}_8$ are taken in the three nucleotides $\text{A}, \text{G}, \text{C}$ (left) and $\text{A}, \text{T}, \text{C}$ (right). The card seq for $\pi_3^{(6)}$ is $[1, 7, 59, 874, 20371, 748320 \cdots]$.

| Triplet | PDB | Conformation | $\pi$ | Triplet | PDB | Conformation | $\pi$ |
|---------|-----|--------------|-------|---------|-----|--------------|-------|
| AGC     | 1ZFM| B            | $F_3$ | ATC     | 1ZFC/1ZF3 | B/J  | $\pi_3^{(6)}$ |
| ACG     | none | unknown      | $F_3$ | ACT     | none | unknown      | $\pi_3^{(3)} \rightarrow \pi_2$ |
| GCA     | 1ZFE| B            | $F_3$ | TCA     | none | unknown      | $\pi_3^{(3)} \rightarrow \pi_2$ |
| GAC     | 1ZF7| B            | $F_3$ | TAC     | none | unknown      | $\pi_3^{(3)} \rightarrow \pi_2$ |
| CAG     | none | unknown      | $F_3$ | CAT     | none | unknown      | $\pi_3^{(3)} \rightarrow \pi_2$ |
| CGA     | none | unknown      | $F_3$ | CTA     | none | unknown      | $\pi_3^{(3)} \rightarrow \pi_2$ |

In Table 6, $\text{N}_6$, $\text{N}_7$ and $\text{N}_8$ are taken in the two nucleotides $\text{A}$ and $\text{T}$, forming eight triplets. The DNA conformation (when known) is of type $\text{B}$. The card seq that we obtain is groups $\pi_3^{(3)}$, $\pi_3^{(4)}$ or $\pi_3^{(5)}$, as described in the caption of Table 5. In six cases over the eight possibilities, the groups encapsulate the topology of the rank 2 group $\pi_2$, whose associated link is shown in Figure 2 (right). As already mentioned, the $\text{SL}(2, \text{C})$ character variety contains the Fricke–Klein seventh variable polynomial $\text{SL}(2, \text{C})$.

In Table 7, $\text{N}_6$, $\text{N}_7$ and $\text{N}_8$ are taken in the two nucleotides $\text{A}, \text{G}$ (left part) and $\text{A}, \text{C}$ (right part). This time, either the card seq of the group $\pi$ is that of the free group $F_3$, of rank 3 (10 cases over the 16 possibilities), or not. In the latter case, the group encapsulates the topology of $\pi_3$ only at the right side of the table. Similar conclusions hold in Table 8 when $\text{N}_6$, $\text{N}_7$ and $\text{N}_8$ are taken in the two nucleotides $\text{A}, \text{G}, \text{C}$ (left part) and $\text{A}, \text{T}, \text{C}$ (right part).

To summarize this section, no clear correlation is observed between the DNA conformations of the considered decamer and our group analysis. Longer sequences may be needed to obtain such a correlation. For instance, the two-letter DNA dodecamer sequence $d(\text{CCCCCGCCGGGG})$ (PDB 2D47) corresponding to a complete turn of A-DNA—see Figure 5 (right)—features the polynomial $f_H$ (for HL) and a fourth-order polynomial similar to $f_R$ with four simple singularities, as announced at the end of Section 3.
Figure 5. (Left) The four-strand Holliday junction $J$: PDB 1ZF2, (Right) A complete turn of A-DNA: PDB 2D47. It is associated to DNA dodecamer sequence $d(CCCCCGCGGGGG)$ with $SL(2, \mathbb{C})$ containing the factor $f_\mathcal{H} = xyz - x^2 - y^2 - z^2 + 4$ (the Cayley cubic) and the factor $f_\overline{\mathcal{H}} = z^4 - 2xyz + 2x^2 + 2y^2 - 3z^2 - 4$.

5. Conclusions

In the present paper, following earlier work about the genetic code [2,3] and about the role of transcription factors in genetics [4], we made use of group theory applied to appropriate DNA motifs and we computed the corresponding variety of $SL(2, \mathbb{C})$ representations. The DNA motifs under consideration may be canonical structures, such as (double-stranded) B-DNA, or non-canonical DNA structures [40], such as (single-stranded) telomeres, (double-stranded) A-DNA, Z-DNA or cruciforms, (triple-stranded) H-DNA, (four-stranded) i-motifs or G-quadruplexes, etc. One objective of the approach is to establish a correspondence between the algebraic geometry and topology of the $SL(2, \mathbb{C})$ character variety and the types of canonical or non-canonical DNA-forms. For example, for two-letter transcription factors, one can correlate the presence of the Cayley cubic and/or a $K_3$ surface in the variety with DNA supercoiling in the remodeling of B-DNA to Z-DNA. For three- or four-letter DNA structures, our work needs to be developed in order to put the features of the variety and potential diseases in correspondence. In a separate work devoted to topological quantum computing, the topology of the four-punctured sphere and the related Fricke surfaces (generalizing the Cayley cubic) are relevant [41]. In addition, Fricke surfaces may be put in parallel with differential equations of the Painlevé VI type. It will be important to compare these models with other qualitative models based on non-linear differential equations [42,43]. In the near future, we intend to apply these mathematical tools to the context of DNA structures.

Author Contributions: Conceptualization, M.P., F.F. and K.I.; methodology, M.P., D.C. and R.A.; software, M.P.; validation, R.A., F.F., D.C. and M.M.A.; formal analysis, M.P. and M.M.A.; investigation, M.P., D.C., F.F. and M.M.A.; writing—original draft preparation, M.P.; writing—review and editing, M.P.; visualization, F.F. and R.A.; supervision, M.P. and K.I.; project administration, K.I.; funding acquisition, K.I. All authors have read and agreed to the published version of the manuscript.

Funding: Funding was obtained from Quantum Gravity Research in Los Angeles, CA, USA.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.
33. Fulnečková, J.; Hasíková, T.; Fajkus J.; Lukešová, A.; Eliáš, M. Sýkorová, E. Dynamic Evolution of Telomeric Sequences in the Green Algal Order Chlamydomonadales. *Genome Biol. Evol.* **2012**, *4*, 248–264. [CrossRef]

34. Gavory, G.; Farrow, M.; Balasubramanian, S. Minimum length requirement of the alignment domain of human telomerase RNA to sustain catalytic activity in vitro. *Nucl. Acids Res.* **2002**, *30*, 4470–4480. [CrossRef]

35. Parkinson, G.N.; Ghosh, R.; Neidle, S. Structural basis for binding of porphyrin to human telomeres. *Biochemistry* **2007**, *46*, 2390–2397. [CrossRef]

36. Zhang, N.; Phan, A.T.; Patel, D.J. (3+1) assembly of three human telomeric repeats into an asymmetric dimeric G-Quadruplex. *J. Am. Chem. Soc.* **2005**, *127*, 17277–17285. [CrossRef]

37. Kang, C.; Zhang, X.; Ratliff, R.; Moyzis, R.; Rich, A. Crystal structure of four-stranded Oxytricha telomeric DNA. *Nature* **1992**, *356*, 126–131. [CrossRef]

38. Laughlan, G.; Murchie, A.I.H.; Norman D.H.; Moore, M.H.; Moody, P.C.E.; Lilley D.M.J.; Luisi, B. The High-Resolution Crystal Structure of a Parallel-Stranded Guanine Tetraplex. *Science* **1994**, *265*, 520–524. [CrossRef]

39. Planat, M.; Aschheim, R.; Amaral, M.M.; Fang, F.; Irwin, K. Quantum information in the protein codes, 3-manifolds and the Kummer surface. *Symmetry* **2021**, *13*, 39. [CrossRef]

40. Bansal, A.; Kaushik, S.; Kukreti, S. Non-canonical DNA structures: Diversity and disease association. *Front. Genet.* **2022**, *13*, 959258. [CrossRef]

41. Planat, M.; Chester, D.; Amaral, M.; Irwin K. Fricke topological qubits. *Preprints*, 2022. [CrossRef]

42. Matsutani, S.; Previato, E. An algebro-geometric model for th shape of supercolied DNA. *Phys. D Nonlinear Phenom.* **2022**, *430*, 133073. [CrossRef]

43. Rand, D.A.; Raju, A.; Saez, M.; Siggia, E.D. Geometry of gene regulatory dynamics. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2109729118. [CrossRef]