Article

DNA sequence and structure under the prism of group theory and algebraic surfaces

Michel Planat [1,*,†] (ID), **Marcelo M. Amaral** [2,†] (ID), **Fang Fang** [2,†] (ID), **David Chester** [2,†] (ID), **Raymond Aschheim** [2,†] (ID) **and Klee Irwin** [2,†] (ID)

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; marcelo@quantumgravityresearch.org (M.M.A.);
    fang@quantumgravityresearch.org (F.F.); davidc@quantumgravityresearch.org (D.C.);
    raymond@quantumgravityresearch.org (R.A.); klee@quantumgravityresearch.org (K.I.)
*   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 π, one can discriminate two important families: (i) the cardinality structure for conjugacy classes of subgroups of π is that of a free group on 1 to 4 bases and the DNA word, viewed as a substitution sequence, is aperiodic. (ii) The cardinality structure for conjugacy classes of subgroups of π is not that of a free group, the sequence is generally not aperiodic and topological properties of π 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 3 distinct bases, most of the time, satisfies (i). For 2-base sequences in the free case (i) or non free case (ii), the topology of π may be found in terms of the $SL(2,\mathbb{C})$ character variety of π and the attached algebraic surfaces. The linking of two unknotted curves –the Hopf link– may occur in the topology of π in cases of biological importance, in telomeres, G-quadruplexes, hairpins and junctions, a feature we already found in the context of models of topological quantum computing. For 3- and 4 base sequences, other knotting configurations are noticed and a building block of the topology is the 4-punctured sphere. Our methods have the potential to discriminate potential diseases associated to the sequences.

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

0. 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 we did in our papers [2,3]. Such an approach of 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 [2, Table 3], [3, Table 4]. 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 which 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 [4, Section 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 1, we briefly account for the many types of topology 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 2, we explain the concept of an $SL(2, \mathbb{C})$ character variety associated to an infinite group with 2 or 3 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 3-dimensional space. In the later case, the variety contains the Fricke-Klein seventh variable polynomial that is characteristic of the topology of the 3-dimensional sphere with four points removed.

In Section 3, we apply these mathematical methods to transcription factors, telomeric sequences and a specific DNA decamer sequence whose almost all conformations have been crystallized.

1. Materials and Methods

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

1.1. DNA conformations
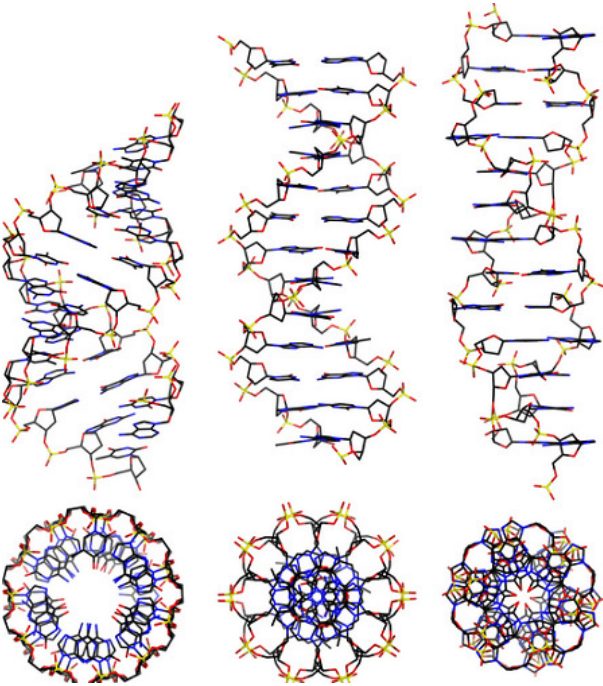
Figure 1. From left to right, the structures of A-, B- and Z-DNA. The view of the double helix from above (or below) shows distinct symmetries, 11-fold for the A-DNA, 10-fold for the B-DNA and 6-fold for the Z-DNA [9,10].

DNA is a long polymer made from a chain of the nucleotides A, T, C, or G. DNA exists in many possible conformations that include a double stranded helix of A-DNA, B-DNA, and Z-DNA, although only B-DNA and Z-DNA has 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 Fig. 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.

1.2. 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 π 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 3 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	$[1,1,1,1,1,1,1,1,1,\cdots]$	A000012
2	$[1,3,7,26,97,624,4163,34470,314493,\cdots]$	A057005
3	$[1,7,41,604,13753,504243,24824785,1598346352,\cdots]$	A057006

Next, given a finitely generated group fp 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. But 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, C and rel is the motif) may not be the free group $F_1 = \langle x_1, x_2 | x_1 x_2 \rangle$, or $F_2 = \langle x_1, x_2, x_3 | x_1 x_2 x_3 \rangle$, or $F_3 = \langle x_1, x_2, x_3, x_4 | x_1 x_2 x_3 x_4 \rangle$. The closeness of f_p to F_r can be checked by its signature in the finite range of indices of the card seq.

1.2.1. 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 [4, Section 4]. Let us introduce the concept of a general substitution rule in the context of free groups. A general substitution rule ρ on a finite alphabet $\mathcal{A}_r$ on r letters is an endomorphism of the corresponding free group F_r [13, Definition 4.1]. The endomorphism property means the two relations $\rho(uv) = \rho(u)\rho(v)$ and $\rho(u^{-1}) = \rho^{-1}(u)$, 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 ρ with the tools of matrix algebra.

The map α 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 ρ, there is a unique mapping $M(\rho)$ that makes the map diagram commutative [13, p. 68]. The substitution matrix $M(\rho)$ of ρ may be specified by its elements at row i and column j as follows

$$(M(\rho))_{i,j} = \text{card}(\rho_{a_i}(a_j)).$$

This approach was applied to binding motifs of transcription factors [4]. The binding motif rel in the finitely presented group $fp = \langle A, T, G, C | rel(A,T,G,C) \rangle$ is split into appropriate segments so that $rel = rel_A rel_T rel_G rel_C$ with the substitution rules $A \to rel_A$, $T \to rel_T$, $G \to rel_G$, $C \to rel_C$.

We are interested in the sequence of finitely generated groups

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

whose card seq is the same at each step l and equal to the card seq of the free group F_r (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.

1.2.2. 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 $M \in \mathrm{Mat}(d, \mathbb{R})$ 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 [13, Corollary 4.3]. A primitive substitution rule ρ of substitution matrix $M(\rho)$ with an irrational PF-eigenvalue is aperiodic.

A well studied primitive substitution rule is The Fibonacci rule $\rho = \rho_F : a \to ab$, $b \to a$ of substitution matrix $M_F = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ and PF-eigenvalue equal to the golden ratio $\lambda_{PF} = \phi = (\sqrt{5} + 1)/2$ [13, Example 4.6]. As expected, the irrationality of λ corresponds to the aperiodicity of the Fibonacci sequence.

The sequence of Fibonacci words is as follows:

$$a, b, ab, aba, abaab, abaababa, abaababaabaab, \cdots$$

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 $\mathrm{rel}(a,b) = ab, aba, abaab, abaababa, \cdots$ have a card seq whose elements are 1's, 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.

1.2.3. 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 $rel = TGAGTCA$ [14]. For this case the 4-letter generated group has card seq similar to the free group F_3 given in Table 1.

We split rel into four segments so that rel $= \mathrm{rel}_A \mathrm{rel}_T \mathrm{rel}_G \mathrm{rel}_C$ with the substitution maps $A \to \mathrm{rel}_A = T$, $T \to \mathrm{rel}_T = G$, $G \to \mathrm{rel}_G = AGTC$, $C \to \mathrm{rel}_C = A$, 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 [137] follows that the selected sequence for the Fos gene is aperiodic. [138]

All 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 [139] freedom correlates to aperiodicity as expected. [140]

Further examples are obtained in the context of DNA sequences for transcription [141] factors [4, Section 4] and below, in relation to DNA conformations and telomeres. [142]

2. Discussion [143]
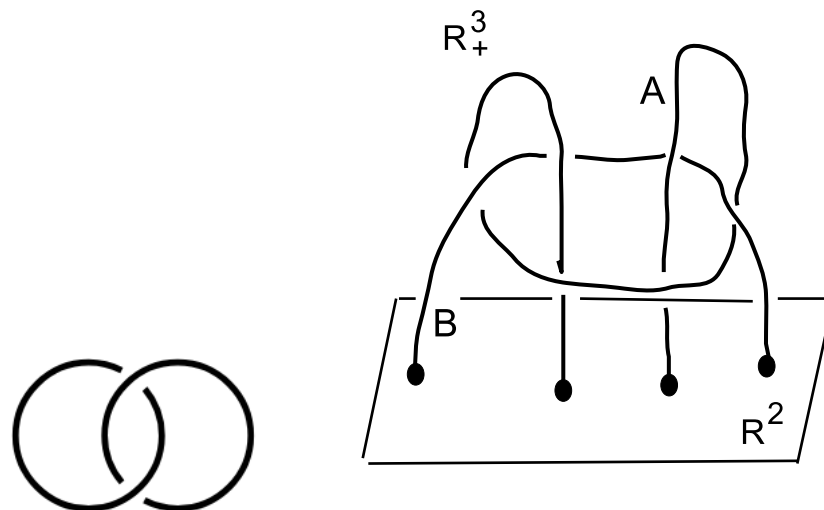
Figure 2. Left: the Hopf link. Right: the link $L = A \cup B$ is attached to the plane R^2 in the half-space R^3_+ . It is not splittable. This can be proved by checking that the fundamental group $\pi = \pi_2(L)$ is not free [15],[16, p. 90]. One gets $\pi_2 = \langle x, y, z | (x, (y, z)) = z \rangle$, where $(.,.)$ means the group theoretical commutator. The cardinality sequence of cc of subgroups of π_2 is $[1, 3, 10, 51, 164, 1365, 9422, 81594, 721305, \cdots]$ [4, Figure 3].

2.1. $SL(2, \mathbb{C})$ character varieties and algebraic surfaces [144]

Recently, we found that the representation theory of finite groups with their character [145] table allows us to derive an approach of the genetic code [3]. [146]

For infinite groups π such as those defined by DNA sequences, it is useful to describe [147] the representations of π in the Lorentz group $SL(2, \mathbb{C})$, the group of (2×2) matrices with [148] complex entries and determinant 1. Such a group expresses the fundamental symmetry of [149] all known physical laws, apart from gravitation. [150]

Representations of π in $SL(2, \mathbb{C})$ are homomorphisms $\rho : \pi \to SL(2, \mathbb{C})$ with character [151] $\kappa_\rho(g) = \text{tr}(\rho(g)), g \in \pi$. The set of characters allows us to define an algebraic set by taking [152]
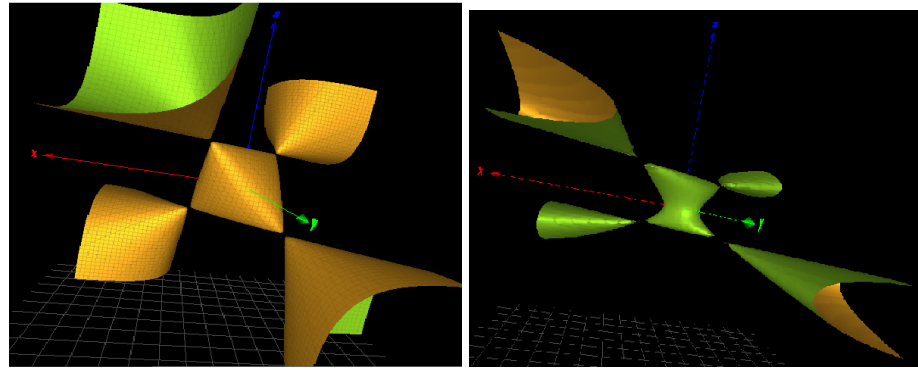
Figure 3. Left: a 3-dimensional picture of the $SL_2(\mathbb{C})$ character variety Σ_H for the Hopf link complement H. Right: a modified character variety of defining equation $f_{\tilde{H}}(x, y, z)($ with similar singularities.

the quotient of the set of representations ρ by the group $SL_2(\mathbb{C})$, which acts by conjugation on representations [17,18].

For two-generator groups, the character variety may be decomposed into the product of surfaces that reveals the topology of 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 [6, Section 2.1] for details.

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

2.2. The Hopf link

Taking the linking of two unknotted curves as in Fig. 3 (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 3-sphere S^3

$$\Pi_1(S^3 \setminus L2a1) = \langle a, b | [a, b] \rangle = \mathbb{Z}^2, \tag{1}$$

where $[a, b] = abAB$ (with $A = a^{-1}$, $B = b^{-1}$) is the group theoretical commutator.

There are interesting properties of the knot group Π_1 of the Hopf link.

First, the number of coverings of degree d of Π_1 (which is also the number of conjugacy classes of index d) is precisely the sum of divisor function $\sigma(d)$ [19].

Second, an invariance of Π_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 [13, Examples 4.5 and 4.6]. Such an observation links the Hopf link, the group Π_1 of the 2-torus and aperiodic substitutions.

Using Sage software [8] developed from Ref. [18], the $SL_2(\mathbb{C})$ character variety is the polynomial corresponding to the so-called Cayley cubic

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

As expected, the 3-dimensional surface $\Sigma : f_H(x, y, z) = 0$ is the trace of the commutator and is known to correspond to the reducible representations [20, Theorem 3.4.1]. A picture is given in Fig. 3 (Left).

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 4 simple singularities.

2.3. 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 Fig. 3 (Right). The defining polynomial is

$$f_{\tilde{H}}(x,y,z) = z^4 - 2xyz \ (+z^3) + 2x^2 + 2y^2 - 3z^2(-4z) - 4. \tag{3}$$

The homogenization of equation $f_{\tilde{H}}(x,y,z)$ allows us to classify it as a conic bundle in the family of K_3 surfaces.

To anticipate the DNA sequence, whose group $\pi_{\tilde{H}}$ leads to the component $f_{\tilde{H}}(x,y,z)$ in the character variety, is d(CCCCGCGGGG) (of A-DNA conformation) with the label 1ZEY in the PDB bank.

The DNA dodecamer sequence d(CCCCCGCGGGGG) is also found in the PDB bank with label 2D47 as corresponding to a comple turn of A-DNA. The character variety for the group defined by this sequence contains the polynomial f_H and a polynomial similar to $f_{\tilde{H}}$ without the third order term z^3 and the first order term $-4z$, but in the same family.

2.4. The Fricke-Klein seventh variable polynomial

The Cayley cubic is a subset of the character variety for the four punctured three-dimensional sphere $\mathbb{S}_4^2$ (the sphere minus 4 points). Its fundamental group Π_1 is isomorphic to the free group F_3 of rank 3, $\Pi_1(\mathbb{S}_4^2) = \langle \alpha, \beta, \gamma, \delta | \alpha\beta\gamma\delta \rangle$, where the four homotopy classes $\alpha, \beta, \gamma, \delta$ correspond to loops around the punctures.

The $SL(2,\mathbb{C})$ character variety for $\Pi_1(\mathbb{S}_4^2)$ satisfies a quartic equation in terms of the Fricke-Klein seventh variable polynomial [20, p 65], [21]

$$f(x,\theta) = xyz + x^2 + y^2 + z^2 - \theta_1 x - \theta_2 x - \theta_3 z + \theta_4, \tag{4}$$

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

3. Results

3.1. Group structure and topology of transcription factors

Table 2. Group structure of motifs for a few 2-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 $[1,3,2,5,2,7,2,8,3,8,2,13,2,9,4,\cdots]$. The card seq for π_1 is $[1,4,1,2,4,2,1,7,2,2,4,2,2,8,1,2,7,2,3,\cdots]$; for π_1' it is $[1,1,1,2,1,3,3,1,2,2,1,1,9,2,14,2,1,\cdots]$. The symbol HL means that the Cayley cubic is part of the Groebner base for the ideal ring of the corresponding $SL(2,\mathbb{C})$ character variety. For 3-letter transcription factors, the ideal ring of the corresponding $SL(2,\mathbb{C})$ character variety contains the Fricke-Klein seventh variable polynomial 4 that is feature of the 4-punctured sphere topology. The group structure of 3-letter transcription factors not leading to free groups is shown in [4, Table 5]

gene	motif	card seq	link	type	Literature
DBX	TTTATTA	F_1	HL	K_3	[22], MA0174.1
SPT15	TATATATAT	.	.	.	.,MA0386.1
PHOX2A	TAATTTAATTA	$\approx F_1$	.	.	., MA0713.1
FOXA	TGTTTGTTT	F_1	.	.	[23,24]
FOXG	TTTGTTTTT	.	.	.	[23]
NKX6-2	TAATTAA	H_3	no	K_3	[22], [MA0675.1, MA0675.2]
FOXG	TGTTTG	$BS(-1,1)$	no	K_3	[25], [22], MA1865.1
HoxA1, HoxA2	TAATTA	π_1	no	K_3	[22], [MA1495.1, MA0900.1]
POU6F1, Vax					., [MAO628.1, MA0722.1]
RUNX1	TGTGGT	.	no	.	., MA0511.1
RUNX1	TGTGGTT	π_1'	no	K_3	[22], MA0002.2
EHF	CCTTCCTC	.	HL		., MA0598.1

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		drosophila segmentation	
SPT15	TATA-box	gene expression, regulation	
	binding protein	in Saccharomyces cerevisiae	
PHOX2A	homeodomain	differentiation, maintenance	fibrosis
		of noradrenergic phenotype	of extraocular muscles
FOX proteins	forkhead box	growth, differentiation,	
FOXA2	.	insulin secretion	diabete
		longevity	
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	.	.	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

In a transcription factor, a motif-specific DNA binding factor controls the rate of 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_1 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 π_1 and π'_1. Compared to 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.

3.1.1. The character variety for the transcription factor of the DBX gene

We explicitly show the $SL(2, \mathbb{C})$ character variety for the transcription factor of the DBX gene.

$$f_{DBX}(x, y, z) = f_H(x, y, z)(yz^2 - y^2 - xz - y + 2)(xy^2 - z^3 - yz - x + 3z)$$
$$(y^3 - z^2 - 3y + 2)(y^2z - xy - yz + x - z)(z^4 - x^2y + xz - 4z^2 + y + 2)$$

$$(5)$$

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 later factor also belongs to the character variety of group $\Pi_1(S^4 \setminus \tilde{E}_6)$, where S_4 is the 4-sphere and $\tilde{E}_6$ is the singular fiber IV^* in Kodaira's classification of minimal elliptic surfaces [6, Fig. 4b].

It is important to mention that, for 3-letter transcription factors, the ideal ring of the corresponding $SL(2, \mathbb{C})$ character variety contains the Fricke-Klein seventh variable polynomial (4) that is a feature of the 4-punctured sphere topology.

Table 3 provides a short account of function or a 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 [26], 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 [27].

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.

Figure 4. Human telomere DNA quadruplex structure in K+ solution hybrid-1 form, PDB 2HY9 [29].

3.2. 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 [28]. 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 $SL(2,\mathbb{C})$ 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 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 Perron-Frobenius eigenvalue λ_{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 π_2, see Fig. 2 (Right) for the definition of such a group. In the former 7 cases, the DNA topology is known to be a G-quadruplex structure [29]-[35]. We could identify an aperiodic structure of the telomere sequence with the Perron-Frobenius eigenvalue λ_{PF} as shown in column 5. In the later case, the topology is of the basket type [35] 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 π 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 Fig. 4. The card seq for π_1'' is $[1, 3, 2, 16, 16, 69, 118, 719, 1877, 8949 \cdots]$ The Hecke group H_4 is defined in [4, Table 2].

seq	organism	PDB	card seq	λ_{PF}	link/DNA conf	ref
G4T4G4	Oxytricha	1D59	π_1''	$(\sqrt{5}+1)/2$	HL	[36]
TG4T	universal	244D_1	H_4	.	no	[37]
T2G4	Tetrahymena	230D	H_4	.	no	[28]
T2AG3	Vertebrates	2HY9	F_2	2.5468	G-quadr.	[29]
TAG3	Giardia	2KOW	F_2	2.2055	G-quadr	[30]
T2AG2	Bombys mori	unknown	F_2	no	G-quadr	[31]
T4AG3	Green algae	unknown	F_2	3.07959	unknown	[32]
G2T2AG	Human	unknown	F_2	2.5468	G-quadr	[33]
TAG3T2AG3	Human	2HRI	F_2	3.3923	G-quadr	[34]
G3T2AG3T2AG3T	Human	unknown	F_2	4.3186	G-quadr	[35]
(GGGTTA)3G3T	Human	unknown	π_2	no	basket	[35]

3.3. Group structure and topology of the DNA decamer sequence $d(CCnnnN_6N_7N_8GG)$[10]

A challenging question of structural biology is to determine if and how a DNA (or RNA) sequence defines the 3-dimensional conformation, as well as the secondary and tertiary structure of proteins. In the previous two subsections, we tackled the problem with regards 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 of the model employed and the choice made to recognize the sections of the secondary structures (e.g. alpha helices, beta sheets and coils) [38]. In the later case, in many organisms, nature invented telomerase for taking care of the replication without damaging too much the sequences at the 3-ends while keeping the catalyzing action of DNA polymerase. Again, there is (a?) loop complex in telomerase comprising telomere-binding proteins, with secondary structures not 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 $d(CCnnnN_6N_7N_8GG)$ the factors N_6, N_7 and N_8 are taken in the two nucleotides G, C and nnn is specified in order to maintain the self-complementarity of the sequence. This inverse repeated motif is the minimum motif 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 4-stranded Holliday junction J. We refer to [10, Table 1] 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 to 8.

In Table 5, N_6, N_7 and N_8 are taken in the two nucleotides G, C forming 8 triplets and the associated 2-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 5 cases over the 9 possibilities, one case (with triplet CGG and code 1ZEY in the protein data bank) shows

Table 5. Group analysis of the sequence $d(CCnnnN_6N_7N_8GG)$, where N_6, N_7 and N_8 are taken in the two nucleotides G, 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 [10, Table 1]), the fourth column is for the cardinality structure of subgroups of π 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 π. The symbols A, B and J are for A-DNA, B-DNA and a 4-stranded Holliday junction, a lower case is used when the conformation is not confirmed in [10].

| triplet	PDB	conformation	card seq (π)	knot
CCC	1ZF1	A	$[1, 1, 1, 1, 7, 1, 1, 2, 9, 6, \cdots]$	HL
CCC	1ZF2	J	idem	HL
CCG	1ZEX	A	idem	HL
CGG	1ZEY	A	$[1, 1, 1, 2, 6, 3, 1, 4, 2, 6, \cdots]$	HL like
CGC	none	unknown	$[1, 1, 2, 1, 6, 3, 2, 1, 3, 6, \cdots]$	no
GGG	1ZF9	A	$[1, 1, 1, 1, 10, 25, 25, 9, 2, 1798, \cdots]$	no
GCC	none	b/J	$[1, 1, 1, 1, 6, 1, 2, 1, 1, 6, \cdots]$	HL
GCG	none	unknown	$[1, 1, 2, 2, 7, 5, 1, 4, 5, 9, \cdots]$	no
GGC	none	B/a	$[1, 1, 1, 1, 6, 11, 9, 5, 2, 208, \cdots]$ (card seq of Hecke group H_5)	no

Table 6. Group analysis of the sequence $d(CCnnnN_6N_7N_8GG)$, where N_6, N_7 and N_8 are taken in the two nucleotides A,T [10]. Groups π_3 and π_3' are as in [4, Table 5]. The card seq for is π_3' it is $[1, 7, 50, 867, 15906, 570528, \cdots]$; for π_3'' it is $[1, 7, 50, 739, 15234, 548439, \cdots]$, for $\pi_3^{(4)}$ it is $[1, 7, 59, 1258, 24787, \cdots]$. Groups π_3'' and π_3' may be simplified to a group whose card seq is that of π_2, the fundamental group of the link $L = A \cup B$ described in Fig. 3 (Right).

triplet	PDB	conformation	π
TTA	1ZFH	B	$\pi_3'' \to \pi_2$
TAA	none	B	$\pi_3'' \to \pi_2$
AAT	none	b	$\pi_3'' \to \pi_2$
ATT	none	unknown	$\pi_3'' \to \pi_2$
AAA	none	b	$\pi_3' \to \pi_2$
TTT	none	unknown	$\pi_3' \to \pi_2$
ATA	none	unknown	$\pi_3^{(4)}$
TAT	none	unknown	$\pi_3^{(4)}$

an algebraic surface similar to the Cayley cubic (with 4 simple singularities) as defined in Equation 3 and as shown in Fig. 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.

In Table 6, N_6, N_7 and N_8 are taken in the two nucleotides A, T forming 8 triplets. The DNA conformation (when known) is of type B. The card seq we obtain is that groups π_3', π_3'' or $\pi_3^{(4)}$ as described in the caption of Table 5. In 6 cases over the 8 possibilities, the groups encapsulate the topology of the rank 2 group π_2 whose associated link is shown in Figure 2 (Right). As already mentioned, the $SL(2, \mathbb{C})$ character variety contains the Fricke-Klein seventh variable polynomial 4.

In Table 7, N_6, N_7 and N_8 are taken in the two nucleotides A, G (left part) and A, C (right part). This time, either the card seq of the group π is that of the free group F_3, of rank 3 (10 cases over the 16 possibilities), or not. In the later case, the group encapsulates the topology of π_2 only at right side of the table. Similar conclusions hold in Table 8 when N_6, N_7 and N_8 are taken in the two nucleotides A, G, C (left part) and A, T, C (right part).
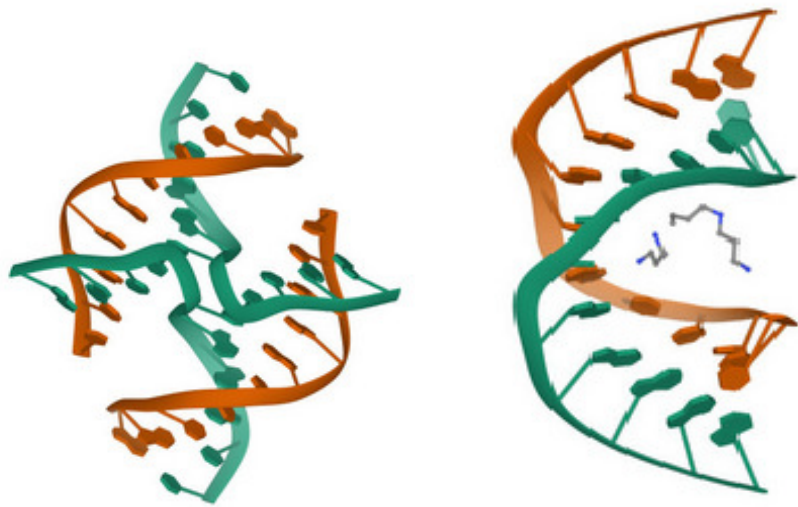
Figure 5. (Left) The 4-strand Holliday junction J: PDB $1ZF2$, (Right) A complete turn of A-DNA: PDB $2D47$. It is associated to DNA dodecamer sequence $d(CCCCCGCGGGGG)$ with $SL(2, \mathbb{C})$ containing the factor $f_H = xyz - x^2 - y^2 - z^2 + 4$ (the Cayley cubic) and the factor $f_{\tilde{H}} = z^4 - 2xyz + 2x^2 + 2y^2 - 3z^2 - 4$.

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 get such a correlation. For instance, the 2-letter DNA dodecamer sequence $d(CCCCCGCGGGGG)$ (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 $f_{\tilde{H}}$ with 4 simple singularities, as announced at the end of Section 2.

Table 7. Group analysis of the sequence $d(CCnnnN_6N_7N_8GG)$[10], where N_6, N_7 and N_8 are taken in the two nucleotides A,G (left) and A,C (right). Groups π_3 and π_3' are as in [4, Table 5]. The card seq for $\pi_3^{(3)}$ it is $[1, 7, 41, 668, 14969, \cdots]$ and for $\pi_3^{(5)}$ it is $[1, 7, 41, 604, 28153, \cdots]$.

triplet	PDB	conformation	π	triplet	PDB	conformation	π
AGA	1ZEW	B	F_3	ACA	none	unknown	$\pi_3^{(3)} \to \pi_2$
AGG	none	unknown	$\pi_3^{(5)}$	ACC	none	J	F_3
GGA	1ZFA	A	F_3	CCA	none	unknown	F_3
AAG	none	unknown	F_3	AAC	1ZF0	B	$\pi_3'' \to \pi_2$
TGT	none	unknown	F_3	TCT	none	b	$\pi_3^{(3)} \to \pi_2$
TGG	1ZF6	A	F_3	TCC	none	unknown	F_3
GGT	1ZF8	A	F_3	CCT	none	b	F_3
TTG	none	unknown	π_3'	TTC	none	B	$\pi_3'' \to \pi_2$

Table 8. Group analysis of the sequence $d(CCnnnN_6N_7N_8GG)$[10], where N_6, N_7 and N_8 are taken in the three nucleotides A ,G, C (left) and A, T, C (right). The card seq for $\pi_3^{(6)}$ is $[1, 7, 59, 874, 20371, 748320 \cdots]$

triplet	PDB	conformation	π	triplet	PDB	conformation	π
AGC	1ZFM	B	F_3	ATC	1ZFC/1ZF3	B/J	$\pi_3^{(6)}$
ACG	none	unknown	F_3	ACT	none	B	$\pi_3^{(3)} \to \pi_2$
GCA	1ZFE	B	F_3	TCA	none	unknown	$\pi_3^{(3)} \to \pi_2$
GAC	1ZF7	B	F_3	TAC	none	unknown	$\pi_3^{(6)}$
CAG	none	unknown	F_3	CAT	none	unknown	F_3
CGA	none	unknown	F_3	CTA	none	unknown	F_3

References

1. Nerlich, B.; Dingwall, R.; Clarke, D. D. The book of life: how the completion of the Human Genome Project was revealed to the public. *health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine* **2002** *6*, 445–469.
2. Planat, M.; Aschheim, R.; Amaral, M. M.; Fang, F.; Irwin, K. Complete quantum information in the DNA genetic code . *Symmetry* **2020** *12*, 1993.
3. Planat, M.; Chester, D.; Aschheim, R.; Amaral, M.M.; Fang, F.; Irwin, K. Finite groups for the Kummer surface: The genetic code and quantum gravity. *Quantum Rep.* **2021** *3*, 68–79.
4. Planat, M.; Amaral, M. M.; Fang F.; Chester, D.; Aschheim, R.; Irwin K. Group theory of syntactical freedom in DNA transcription and genome decoding. *Curr. Issues Mol. Biol.* **2022** *44*, 1417–1433.
5. Irwin, K. The code-theoretic axiom; the third ontology, Rep. Adv. Phys. Sci **2019** *3*, 1950002.
6. Planat, M.; Amaral, M. M.; Fang F.; Chester, D.; Aschheim, R.; Irwin K. Character varieties and algebraic surfaces for the topology of quantum computing. *Symmetry* **2022** *14*, 915.
7. Bosma, W.; Cannon, J. J.; Fieker, C. ; Steel, A. (eds). *Handbook of Magma functions*, Edition 2.23 (2017), 5914pp (accessed on 1 January 2022).
8. Python code to compute character varieties. Available online: http://math.gmu.edu/s̄lawton3/Main.sagews (accessed on 1 May 2021).
9. DNA. Available online: https://en.wikipedia.org/wiki/DNA (accessed on 1 January 2022).
10. Hays, F. A.; Teegarden, A.; Jones Z. J.R.; Harms, M.; Raup, D.; Watson, J.; Cavaliere, E., Ho, P. S. How sequence defines structure: a crystallographic map of DNA structure and conformation. *PNAS* **2005** *20*, 7157-7162.
11. J. H. Kwak and R. Nedela, Graphs and their coverings. *Lect. Notes Ser.* **2007** *17*, 118.
12. The on-line encyclopedia of integer sequences, available online: https://oeis.org/book.html (accessed on 1 June 2022).
13. Baake M., Grimm U. *Aperiodic order, Vol. I: A mathematical Invitation*; Cambridge Univ. Press: Cambridge, UK, 2013.
14. Glover, J. N.; Harrison, S. C. Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature* **1995** *373*, 257–261.
15. Zeeman E. C. Linking spheres. *Abh. Math. Sem. Univ. Hamburg* **1960** *24*, 149–153.
16. Rolfsen D. *Knots and Links*; AMS Chelsea Publishing: Providence, Rhode Island, USA, 2000.
17. Culler, M.; Shalen P. B. Varieties of group representations and splitting of 3-manifolds, *Ann. of Math.* **1983**, *117*, 109–146.
18. Ashley, C.; Burelle J. P.; Lawton, S. Rank 1 character varieties of finitely presented groups, *Geom. Dedicata* **2018** *192*, 1–19.
19. Liskovets, V.; Mednykh, A. On the number of connected and disconnected coverings over a manifold. *Ars Math. Contemp.* **2009** *2*, 181–189.
20. Goldman, W. M. Trace coordinates on Fricke spaces of some simple hyperbolic surfaces. In *Handbook of Teichmüller theory*, Eur. Math. Soc., Zürich, **2009** *13*, 611-684.
21. Cantat, S.; Loray, F. Holomorphic dynamics, Painlevé VI equation and character varieties. arXiv **2007**, arXiv:0711.1579v2 [math.DS].
22. Sandelin, A.; Alkema, W; Engström, P; Wasserman, WW; Lenhard, B, JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research* **2004** *32*, D91–D94; software available at https://jaspar.genereg.net/
23. Lambert S. A.; Jolma A.; Campitelli L. F.; Das P. K.; Yin Y.; Albu M.; Chen X.; Talpale J.; Hughes T. R.; Weirauch M. T., The human transcription factors. *Cell* **2018** *172*, 650–665. The classification is available at http://www.edgar-wingender.de/huTF_classification.html.

24. Lantz, K. A.; Vatamaniuk, M. Z.; Brestelli, J. E.; Friedman, J. R.; Matschinsky, F. M.; Kaestner, K. H. Foxa2 regulates multiple pathways of insulin secretion. *J. Clin. Invest.* **2004** *114*, 512–520.

25. José-Edwards, D. S; Oda-Ishii, I.; Kugler J. E., Passamaneck, Y. J.; Katikala, L.; Nibu, Y.; Di Gregorio, A. Brachyury, Foxa2 and the cis-Regulatory Origins of the Notochord. *PLoS Genet.* **2015** *11*, e1005730. doi: 10.1371/journal.pgen.1005730.

26. Ray, B. K.; Dhar, S.; Shakya, Ray, A. Z-DNA-forming silencer in the first exon regulates human ADAM-12 gene expression. *PNAS* **2011** *108*, 103–106.

27. Ravichandran, S.; Subramani, V. K.; Kil, K. K. Z-DNA in the genome: from structure to disease. *Biophys. Rev.* **2019** *11*, 383–387.

28. Nugent, C. I; Lundblad, V. The telomerase reverse transcriptase: components and regulation. *Genes & Developement* **1998** *12*, 1073–1085.

29. Dai, J.; Punchihewa, C.; Ambrus, A.; Chen D.; Jones R. A.; Yang, D. Structure of the intramolecular human telomeric G-quadruplex in potassium solution: a novel adenine triple formation. *Nucl. Acids Res.* **2007** *35*, 2440–2450.

30. Hu, L.; Lim, K. W.; Bouazuz, S.; Phan, A. T. Giardia Telomeric Sequence d(TAGGG)4 Forms Two Intramolecular G-Quadruplexes in K+ Solution: Effect of Loop Length and Sequence on the Folding Topology. *J. Am. Chem. Soc.* **2009** *131*, 16824–16831.

31. Kettani, A.; Bouazziz, S.; Wang, W.; Jones, R. A.; Patel, D. J. Bombyx mori single repeat telomeric DNA sequence forms a G-quadruplex capped by base triads. *Nat. Struc. Biol.* **1997** *4*, 383–389.

32. 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.

33. 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.

34. Parkinson, G. N.; Ghosh, R.; Neidle, S. Structural basis for binding of porphyrin to human telomeres. *Biochemistry* **2007** *46*, 2390-2397.

35. 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.

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

37. 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.

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