Article

Algebraic morphology of DNA–RNA transcription and regulation

Michel Planat [1,*,†], Marcelo M. Amaral [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; marcelo@quantumgravityresearch.org (M.M.A.); klee@quantumgravityresearch.org (K.I.)
* Correspondence: michel.planat@femto-st.fr
† These authors contributed equally to this work.

Abstract: Transcription factors (TFs) and microRNAs (miRNAs) are co-actors in genome-scale decoding and regulatory networks, often targeting common genes. In this paper, we describe the algebraic geometry of both TFs and miRNAs thanks to group theory. In TFs, the generator of the group is a DNA-binding domain while, in miRNAs, the generator is the seed of the sequence. For such a generated (infinite) group π, we compute the $SL(2,\mathbb{C})$ character variety, where $SL(2,\mathbb{C})$ is simultaneously a 'space-time' (a Lorentz group) and a 'quantum' (a spin) group. A noteworthy result of our approach is to recognize that optimal regulation occurs when π looks like a free group F_r ($r = 1$ to 3) in the cardinality sequence of its subgroups, a result obtained in our previous papers. A non free group structure features a potential disease. A second noteworthy result is about the structure of the Groebner basis $\mathcal{G}$ of the variety. A surface with simple singularities (like the well known Cayley cubic) within $\mathcal{G}$ is a signature of a potential disease even when G looks like a free group F_r in its structure of subgroups. Our methods apply to groups with a generating sequence made of two to four distinct DNA/RNA bases in $\{A,T/U,G,C\}$. Several human TFs and miRNAs are investigated in detail thanks to our approach.

Keywords: Transcription factors; micro-RNAs; diseases; finitely generated group; $SL(2,\mathbb{C})$ character variety, algebraic surfaces

0. Introduction

Recently, we wrote a paper about a common algebra possibly ruling the beauty and structure in poems, music and proteins [1]. We found that free groups govern the structure of such disparate topics where a language emerges from pure randomness. We coined the concept of 'syntactical freedom' for qualifying this occurrence of symbols organized according to aperiodicity [2,3]. According to our view, the escape to 'syntactical freedom' means a lack of beauty and the signature of a potential disease at the genome scale. The secondary structures in the sequence of proteins or viruses are, most of the time, organized according to the rules of free groups and, otherwise they may be a witness of a potential aberrant topology. One favorite decomposition of the secondary structure of proteins is in term of α-helices, β-sheets and coils [4] but this decomposition and the resulting syntax is model dependent [1].

Apart from the canonical double helix B-DNA we now know that there exists a diversity of non-canonical coding/decoding sequences organized in structures such as Z-DNA (often encountered in transcription factors), G-quadruplex (in telomeres) and other types that are single-stranded, two-stranded or multistranded [5]. RNA is usually a single-stranded molecule in a short chain of nucleotides, as is the case for a messenger RNA (mRNA) or a (non-coding) microRNA (miRNA).

In our approach, the investigated sequences define a finitely generated group f_p whose structure of subgroups is close or away from a free group F_r of rank r, where $r+1$ is

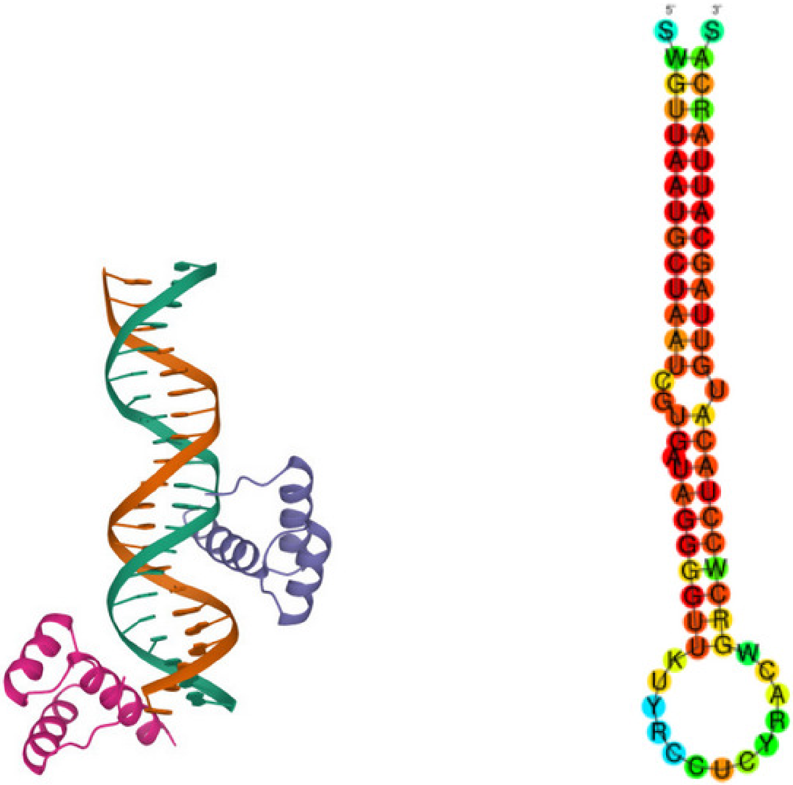

Figure 1. Left: the Nanog transcription factor (PDB 9ANT). Right: the pre-miR-155 secondary structure [10].

the number of distinct amino acids in the sequence (or the number of distinct secondary structures considered in the protein chain). Recently, we also introduced concepts for representing the groups f_p over the Lie group $SL(2,\mathbb{C})$. The $SL(2,\mathbb{C})$ character variety of f_p and its Groebner basis are topological ingredients, they feature algebraic geometric properties of the group f_p under question [6,7]. For the definition of the Groebner basis of an ideal containing multivariate polynomial rings, the reader may consult Reference [8].

In this paper, we will focus upon transcription factors and miRNAs, both serve at properly decoding and regulating the genes and their action, either independently of each other or together by targeting common genes [9]. Figure 1 (Left) is a picture of the pluripotent transcription factor Nanog. Figure 1 (Right) is an example of a pre-miRNA associated to a disease [10]. Both will be investigated in detail in this paper.

In Section 1, we describe the mathematical methods and the software needed for describing the algebraic surfaces relevant to DNA/RNA sequences. This includes the definition of infinite groups under question, of the free groups F_r of rank $r = 1$ to 3 corresponding to 2- to 4-base sequences and the calculation of $SL(2,\mathbb{C})$ representation of such groups. A special care is needed to compute a Groebner basis of the character variety.

Section 2 is a discussion about the type of singular surfaces encountered in this research. They play an imprtant role in our view of a potential disease.

In Section 3, the methods are applied to representative examples of sequences taken from transcription factors and microRNAs whose group is close or away from a free group, and whose Groebner basis of the variety contains simple singularities. Figures 2 to 6 feature the main topological ingredients resulting from this research.

1. Materials and Methods

1.1. Finitely generated groups, free groups and their conjugacy classes

A 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 [3,11]. In the following, the cardinality structure of conjugacy classes of index d in F_r is called the 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/RNA sequence. The card seq of F_r is in Table 1 for the 3 sequences of interest in the context of DNA/RNA.

Table 1. Number of conjugacy classes of subgroups of index d in free group of rank r = 1 to 3. 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 cardinality sequence (card seq) of its conjugacy classes. Often, the DNA/RNA 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/U, 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. The $SL(2,\mathbb{C})$ character variety of a finitely generated group and a Groebner basis

Let f_p be a finitely generated group, we describe the representations of f_p in the (double cover of the) Lorentz group $SL(2,\mathbb{C})$, the group of (2×2) matrices with complex entries and determinant 1. The group $SL(2,\mathbb{C})$ may be seen simultaneously as a 'space-time' (a Lorentz group) and a 'quantum' (a spin) group.

Such a group contains representations as degrees of freedom for all quantum fields and is the gauge group for Einstein-Cartan theory which contains the Einstein-Hilbert action and Einstein's field equations [13]. The Holst action used in loop quantum gravity has quantum gravity states given in terms of $SL(2,\mathbb{C})$ representations [14].

Representations of f_p in $SL(2,\mathbb{C})$ are homomorphisms $\rho : f_p \to SL(2,\mathbb{C})$ with character $\kappa_\rho(g) = \mathrm{tr}(\rho(g))$, $g \in f_p$. The set of characters allows to define an algebraic set by taking the quotient of the set of representations ρ by the group $SL(2,\mathbb{C})$, which acts by conjugation on representations [15,16].

For the effective calculations of the character variety, we make use of a software on Sage [17]. We also need Magma [18] for the calculation of a Groebner basis, at least for 3- and 4-base sequences.

1.3. Algebraic geometry and topology of DNA/RNA sequences

1.3.1. Two-base sequences

Following [19], in this section, we describe the special case of representations for the punctured torus $S_{1,1}$ and the relevance of the extended mapping class group $\mathrm{Mod}^{\pm}(S_{1,1})$ in its action on surfaces of type $\kappa_d(x,y,z)$, $d \in \mathbb{C}$.

Let us take the example of the punctured torus $T_{1,1}$ whose fundamental group, that we denote π, is the free group $F_2 = \langle a, b | \emptyset\rangle$ on two generators a and b. The boundary component of $T_{1,1}$ is a single loop around the puncture expressed by the commutator $[a,b] = abAB$ with $A = a^{-1}$ and $B = b^{-1}$. We introduce the traces

$$x = \mathrm{tr}(\rho(a)),\ y = \mathrm{tr}(\rho(b)),\ z = \mathrm{tr}(\rho(ab)).$$

The trace of the commutator is the surface [15,19]

$$\mathrm{tr}([a,b]) = \kappa_2(x,y,z) = x^2 + y^2 + z^2 - xyz - 2.$$

Another noticeable surface is obtained from the character variety attached to the fundamental group of the Hopf link $L2a1$ that links two unknotted curves. For the Hopf link, the fundamental group is

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

and the corresponding character variety is the Cayley cubic [6]

$$\kappa_4(x,y,z) = x^2 + y^2 + z^2 - xyz - 4.$$

Surfaces κ_2 and κ_4 have been obtained from two different mathematical concepts, from topological and algebraic concepts in dimension 2, respectively. To relate them one makes use of the Dehn-Nielsen-Baer theorem applied to the once punctured torus [20]. According to this theorem, for a surface of genus $g \geq 1$, we have

$$\mathrm{Mod}^{\pm}(S_g) \cong \mathrm{Out}(\pi(S_g)),$$

where the mapping class group $\mathrm{Mod}(S)$ denotes the group of isotopy classes of orientation-preserving diffeomorphisms of S (that restrict to the identity on the boundary ∂S if $\partial S \neq \emptyset$), the extended mapping class group $\mathrm{Mod}^{\pm}(S)$ denotes the group of isotopy classes of all homeomorphisms of S (including the orientation-reversing ones) and $\mathrm{Out}(\pi)$ This leads to the (topological) action of $\mathrm{Mod}^{\pm}$ on the punctured torus as follows

$$\mathrm{Mod}^{\pm}(S_{1,1}) = \mathrm{Out}(F_2) = GL(2,\mathbb{Z}). \tag{1}$$

The automorphism group $\mathrm{Aut}(F_2)$ acts by composition on the representations ρ and induces an action of the extended mapping class group $\mathrm{Mod}^{\pm}$ on the character variety by polynomial diffeomorphisms of the surface κ_d defined by [19]

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

The Cayley surface $\kappa_4(x,y,z)$ possesses 4 simple singularities. We already showed that it plays a role in the context of Z-DNA conformations of transcription factors [7, Tables 2 and 5]. See also the section 3 below and notably Figures 3 (Left) and 6 (Left).

The surface $\kappa_3(x,y,z)$ lies within the character variety for the fundamental group of the link L6a1 [21]. We show below that this surface also lies in the generic Groebner basis obtained for 4-base sequences, see Figure 6 (Right) below.

1.3.2. Three-base sequences

Our main object in this section is the four punctured sphere for which the fundamental group is the free group F_3 of rank 3 whose character variety generalizes the Fricke cubic surface (2) to the hypersurface $V_{a,b,c,d}(\mathbb{C})$ in $\mathbb{C}^7$.

We follow the work of references [15,19,22].

Let $S_{4,2}$ be the quadruply-punctured sphere. The fundamental group for $S_{4,2}$ can be expressed in terms of the boundary components A, B, C, D as $\pi(S_{4,2}) = \langle A,B,C,D|ABCD\rangle \cong F_3$.

A representation $\pi \to SL(2,\mathbb{C})$ is a quadruple

$$\alpha = \rho(A),\ \beta = \rho(B)\ \gamma = \rho(C),\ \delta = \rho(D) \in SL(2,\mathbb{C}) \text{ where } \alpha\beta\gamma\delta = I.$$

Let us associate the seven traces

$$a = \operatorname{tr}(\rho(\alpha)),\ b = \operatorname{tr}(\rho(\beta)),\ c = \operatorname{tr}(\rho(\gamma)),\ d = \operatorname{tr}(\rho(\delta))$$
$$x = \operatorname{tr}(\rho(\alpha\beta)),\ y = \operatorname{tr}(\rho(\beta\gamma)),\ z = \operatorname{tr}(\rho(\gamma\alpha)),$$

where a, b, c, d are boundary traces and x, y, z are traces of elements AB, BC, CA representing simple loops on $S_{4,2}$.

The character variety for $S_{4,2}$ satisfies the equation [15, Section 5.2],[19, Section 2.1],[22, Section 3B], [23, Eq. 1.9] or [24, Eq. (39)]

$$V_{a,b,c,d}(x,y,z) = x^2 + y^2 + z^2 + xyz - \theta_1 x - \theta_2 y - \theta_3 z - \theta_4 = 0 \tag{3}$$

with $\theta_1 = ab + cd$, $\theta_2 = ad + bc$, $\theta_3 = ac + bd$ and $\theta_4 = 4 - a^2 - b^2 - c^2 - d^2 - abcd$.

The 4-punctured sphere, whose fundamental group is the free group F_3 with generator the product of the 4 letters, is a generic topology. It is straightforward to check that the Groebner basis for F_3 contains (among other surfaces and depending on the choice of parameters) a single copy of the generic surfaces $\kappa_4(x,y,z)$, $\kappa_3(x,y,z)$ and $V_{1,1,1,1}(x,y,z) = xyz + x^2 + y^2 + z^2 - 2x - 2y - 2z + 1$, a surface we also denote $f^{(3A_1)}(x,y,z)$ because it contains 3 simple singularities of type A_1 as shown in Figures 2 and 6 (Right).

There are other surfaces encountered in our study of the Groebner basis for transcription factors and miRNAs when the generated group is close or away from the free group F_2 (for 3-base sequences) or the free group F_3 (for 4 base sequences). These surfaces are described in Section 3.

1.3.3. Four-base sequences

There does not exist a huge difference in the structure of a Groebner basis of the character variety in the case of a 4-base sequence compared to the case of a 3-basis sequence. One difference is that one has to manage a 14-dimensional hypersurface $V_{a,b,c,d,e,f,g,h}(x,y,z,u,v,w)$ in $\mathbb{C}^{14}$ (instead of a 7-dimensional one as in the previous subsection). In general, after the appropriate choice of the 8 parameters a,b,c,d,e,f,g,h, the Groebner basis contains more than one copy of the generic Groebner basis, as shown in Table 4. Each copy S of a relevant surface may be of the form $S(x,y,z)$, $S(x,u,v)$, $S(y,u,w)$ or $S(z,v,w)$.

2. Discussion

Given an ordinary projective surface S in the projective space P^3 over a number field, if S is birationally equivalent to a rational surface, the software Magma [18] determines the map to such a rational surface and returns its type within five categories. The returned type of S is P^2 for the projective plane, a quadric surface (for a degree 2 surface in P^3), a rational ruled surface, a conic bundle or a degree p Del Pezzo surface where $1 \le p \le 9$.

A further classification may be obtained for S in P^3 if S has at most point singularities. Magma computes the type of S (or rather, the type of the non-singular projective surfaces in its birational equivalence class) according to the classification of Kodaira and Enriques [25]. The first returned value is the Kodaira dimension of S, which is $-\infty$, 0, 1 or 2. The second returned value further specifies the type within the Kodaira dimension $-\infty$ or 0 cases (and is irrelevant in the other two cases).

Kodaira dimension $-\infty$ corresponds to birationally ruled surfaces. The second return in this case is the irregularity $q \ge 0$ of S. So S is birationally equivalent to a ruled surface over a smooth curve of genus q and is a rational surface if and only if q is zero.

Kodaira dimension 0 corresponds to surfaces which are birationally equivalent to a K_3 surface, an Enriques surface, a torus or a bi-elliptic surface.

Every surface of Kodaira dimension 1 is an elliptic surface (or a quasi-elliptic surface in characteristics 2 or 3), but the converse is not true: an elliptic surface can have Kodaira dimension $-\infty$, 0 or 1.

Surfaces of Kodaira dimension 2 are algebraic surfaces of general type.

One important attribute of a surface is its degree of singularity. Most surfaces S of interest below are almost not singular in the sense that they have at worst simple singularities. The type and the number of simple singularities are denoted in an exponent such as $S^{(lA_1)}$ for l singularities of type A_1. The notation A_1 refers to the type of a Coxeter root system. No other types of singular surfaces are encountered in our paper. All such surface are degree 3 del Pezzo surfaces. For the Cayley cubic $f^{(4)}(x,y,z)$, the exponent (4) means $(4A_1)$. Of course there are non singular surfaces in the character variety for some sequences, corresponding to the other types. But we do not discuss them in this paper.

There are additional facilities offered by Magma for studying a singular scheme. A scheme in Magma is any geometric object defined by the vanishing of polynomials in a projective space. An algebraic surface is a scheme. If the scheme is singular, one can calculate its singular subscheme, its degree and its support, that are important signatures of the scheme. Most of the time, in the examples of this paper, for a singular surface $S^{(lA1)}$, the degree is l and the support contains l simple singular points. Otherwise, we add a lower index to $S^{(lA_1)}$ to qualify the index and the support of the singular subscheme. The notation $S^{(lA1)}_{m,\{\}}$ means that there are l singularities of type A_1, that the degree of the singular subscheme is m and that the support is the empty set.

By the way, transcription factors and microRNAs seem to have smooth rules underlying the singularities of their sequences. This contrasts with some sequences encountered in another context. In [7], we found two cases of two-base sequences whose corresponding Groebner basis contains the surfaces

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

A plot of one of them in [7, Figure 3 (Right)] shows that they look like a generalization $f_{\tilde{H}}(x,y,z)$ of the Cayley cubic $f_H^{(4)}(x,y,z)$. These surfaces are conic bundles in the family of K_3 surfaces. In contrast to $f_H^{(4)}(x,y,z)$, according to Magma, the resolution of such singular surfaces $f_{\tilde{H}}(x,y,z)$ leads to many non isolated singularities (one cannot desingularise the surfaces by blow up).

To summarize the important issues below, a noticeable result of our approach is to recognize that optimal regulation occurs when the group underlying the sequence looks like a free group F_r ($r = 1$ to 3) in the cardinality sequence of its subgroups, a result obtained in our previous papers. A non free group structure features a potential disease. A second noticeable result is about the structure of the Groebner basis of the variety. A surface with simple singularities (like the well known Cayley cubic) within the Groebner basis is a signature of a potential disease even when the generated group looks like a free group F_r in its structure of subgroups. Our methods apply to groups with a generating sequence made of two to four distinct DNA/RNA bases in $\{A, T/U, G, C\}$. Several human TFs and miRNAs are investigated in detail thanks to our approach.

3. Results

In this section, we apply the $SL(2,\mathbb{C})$ representation theory to groups generated by DNA/RNA sequences occurring in transcription factors and microRNAs. Both play a leading role in the decoding of the genome and in genome-scale regulatory networks. Two-letter transcription factors (TFs) whose structure is close or away from the free group F_1 were already investigated in [7, Table 2]. The occurrence of the Cayley cubic $\kappa_4(x,y,z)$ in the Groebner basis of the character variety was found to be a signature in the former case. In this case this surface seems to possess a regulatory action that may be lost in the latter case. In [7, Table 3], the potential diseases associated to a non-free group structure are mentioned. In the present section, one explores in Table 2 the case of a 3-letter sequence of a TF and in Table 3 the case of 3-letter sequence of a miRNA. The case of 4-letter sequence of a miRNA is summarized in Table 4. The role of surfaces with simple singularities in the Groebner basis is emphasized.

3.1. *Algebraic morphology of the transcription factor Prdm1*

The transcriptional repressor PR domain containing 1 (Prdm1), also known as B-lymphocyte-induced maturation protein-1 (Blimp1), is essential for normal development and immunity [26]. It is of a zinc finger type. The consensus sequence ACTTTC corresponds to the code MA0508.2 in [27].

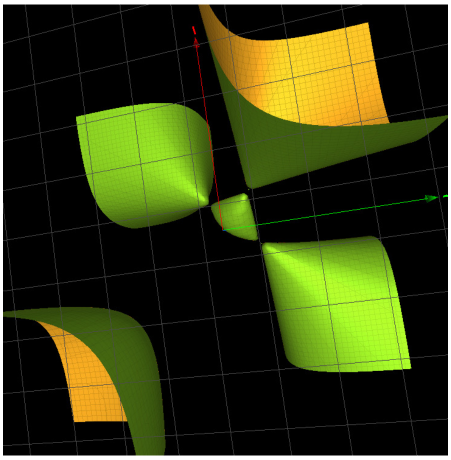

Figure 2. The Fricke surface $V_{1,1,1,1}(x,y,z) = f_a^{(3A_1)}(x,y,z)$ (with three simple singularities of type A_1).

The character variety

The ideal for the character variety $f_{\text{Prdm1}}(a,b,c,d)(x,y,z)$ for a few values of the parameters is

$$f_{\text{Prdm1}}(0,0,0,0) = \kappa_{-4}(x,y,z)(yz+x+2),$$

$$f_{\text{Prdm1}}(0,1,1,0) = y\kappa_{-2}(x,y,z)(x-1),$$

$$f_{\text{Prdm1}}(0,1,0,0) = z\kappa_{-3}(x,y,z)(z^2+1)(yz+x+1)(yz+x+2),$$

$$f_{\text{Prdm1}}(1,1,1,1) = f_a^{(3A_1)}(x,y,z)(y+1)(y+z-1),$$

where $\kappa_{-2}(x,y,z)$, $\kappa_{-3}(x,y,z)$ are Fricke surfaces [21] and $f_a^{(3A_1)}(x,y,z) = xyz + x^2 + y^2 + z^2 - 2x - 2y - 2z + 1$ is the surface drawn in Figure 2. The subscript $3A_1$ is for featuring the three singularities of type A_1.

The Groebner basis

The singular surfaces found in the Groebner basis of the ideal are not similar to those in the ideal. One of them $S_1 = S^{(A_1A_3)}_{3,\{0:1:0:1\}} = 2yz^2 + x^2 + 3z^2 - 2xz - 2yz - 2y - 2z - 2$ is obtained at values $(a,b,c,d) = (1,1,1,1)$ featuring two simple singularities of type A_1 and A_3 with a singular subscheme of degree 3 and the singular point of type A_3 in its support. The other surface obtained in the Groebner basis at values $(a,b,c,d) = (0,0,0,0)$ is a conic bundle of the K_3 type $S_2 := z^4 + 2yz^3 + x^2 - 6yz - 2x - 8$ whose singular subscheme is non zero dimensional and of degree 1.

These two surfaces are non standard in our context of TFs and miRNAs.

3.2. *Algebraic morphology of homeodomains for Nanog and Xvent*

The pluripotency in embryonic stem cells and their regulation is characterized by the expression of several transcription factors [28,29]. Among them, the transcription factor Nanog is present in the embryonic stages of life of several vertebrate species. Nanog binds to promoter elements of hundreds of target genes as a regulatory element. It has a conserved DNA-binding homeodomain with consensus sequence TAATGG. The closest homolog of Nanog is the (nonmammalia) Xenopus, a Xvent transcription factor with consensus sequence CTAATT [29]. In this subsection, we investigate the algebraic morphology of both transcription factors Nanog and Xvent thanks to their consensus sequences.

Table 2. A few (three-base) transcription factors whose group structure is away from a free group or whose Groebner basis of the $SL(2,\mathbb{C})$ character variety contains a (possibly almost) singular surface. The symbol gene is for the identification of the transcription factor in the Jaspar database [27], motif is for the consensus sequence of the transcription factor, card seq is for the cardinality sequence of conjugacy classes of subgroups of the group whose motif is the generator, sing is for the identification of a singular surface within the Groebner basis, the last columnn is for a reference paper and the corresponding disease. The group F_2 is the free group of rank two. The card seq for π_2 is $[1,3,10,51,164,1230,7829,59835,491145\cdots]$, close to the card seq of the group $\langle x,y,z|(x,(y,z))=z\rangle$. The later group is found as governing the structure of many transcription factors and is associated to the link found in [7, Figure 2]. The card seq for π_3 is $[7,14,89,264,1987,11086,93086\cdots]$. The surface $f_b^{(A_1)}(x,y,z) = x^2+y^2-6z^2+4xyz$ (not defined in the text) is part of the character variety for the genes Pitx1, OTX1, ...

gene	motif	card seq	sing	ref & disease
Prdm1	ACTTTC	F_2	$S_1, S_2(x,y,z)$	[27],MA0508.2 lupus, rheumatoid arthritis
POU6F1	TAATGAG	π_2	no sing	.,MA1549.1 lung adenocarcinoma
ELK4	CTTCCGG	.	no sing, Fricke	.,MA0076.2 gastric cancer
OTX2	GGATTA	π_3	no sing	.,[MA0712.2, MA0883.1] medulloblastomas
N-box	TTCCGG	.	no sing, Fricke	[30] drug sensitivity
Pitx1,OTX1,$\cdots$	TAATCC	.	$f_H^{(4)}, f_b^{(A_1)}(x,y,z)$	[27],[MA0682.1,MA0711.1] autism, epilepsy, $\cdots$
Nanog	TAATGG	.	$f_H^{(4)}, f_a^{(A_1)}(x,y,z)$	[28] cancer cells
Xvent	CTAATT	F2	$f_{4,\{\}}^{(2A_1)}, f^{(A_2)}(x,y,z)$	[29]

The Groebner basis for Xvent $f_{\text{Nanog}}(0,0,0,0)$ takes the form

$$f_{\text{Nanog}}(0,0,0,0) = f_H^{(4)}(x,y,z) f_a^{(A_1)}(x,y,z)\cdots$$

where $f_H^{(4)}(x,y,z)$ is the Cayley cubic (with its 4 simple singularities) and $f_a^{(A_1)}(x,y,z) = x^2+y^2-z^2+xyz$ (a surface with a single simple singularity of type A_1) as shown in Figure 3 (Right). The forgotten factors are factors for planes or trivial smooth surfaces.

The Groebner basis for Xvent $f_{\text{Xvent}}(1,1,1,1)$ takes the form

$$f_{\text{Xvent}}(1,1,1,1) = f_b^{(3A_1)}(x,y,z)\cdots$$

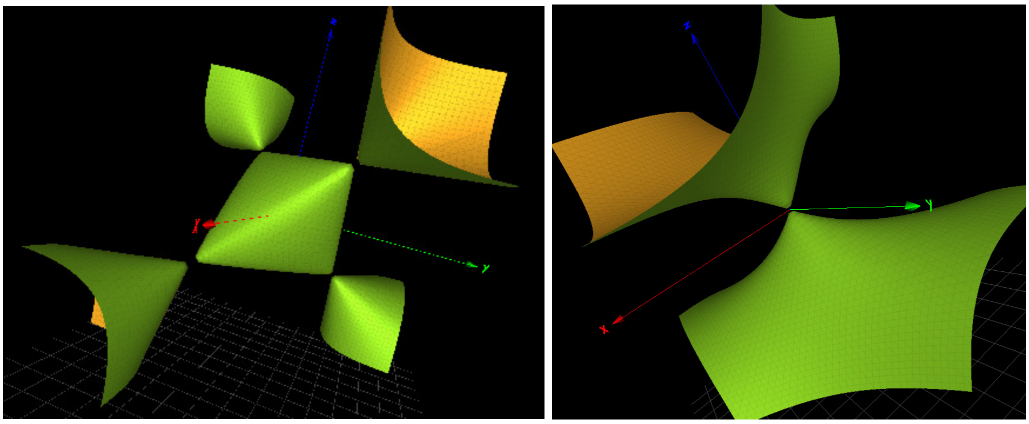

Figure 3. Left: the Cayley cubic $\kappa_4(x,y,z)$. Right: the surface $f_a^{(A_1)}(x,y,z)$.

where $f_b^{(3A_1)}(x,y,z) = x^2 + y^2 + xyz - xy - z - 1$ (a surface with three simple singularity of type A_1). The missing term does not contain surfaces with singularities. The character variety $f_{\mathrm{Xvent}}(0,0,0,0)$ contains the cubic surface $f_{4,\{\}}^{(2A_1)}(x,y,z) = 2z^3 + x^2z + 2xyz + 2y^2 - z^2 - 6z$ (with two simple singularities of type A_1) and other factors for planes or trivial smooth surfaces. Both surfaces $f_b^{(3A_1)}(x,y,z)$ and $f_{4,\{\}}^{(2A_1)}(x,y,z)$ are pictured in Figure 4.

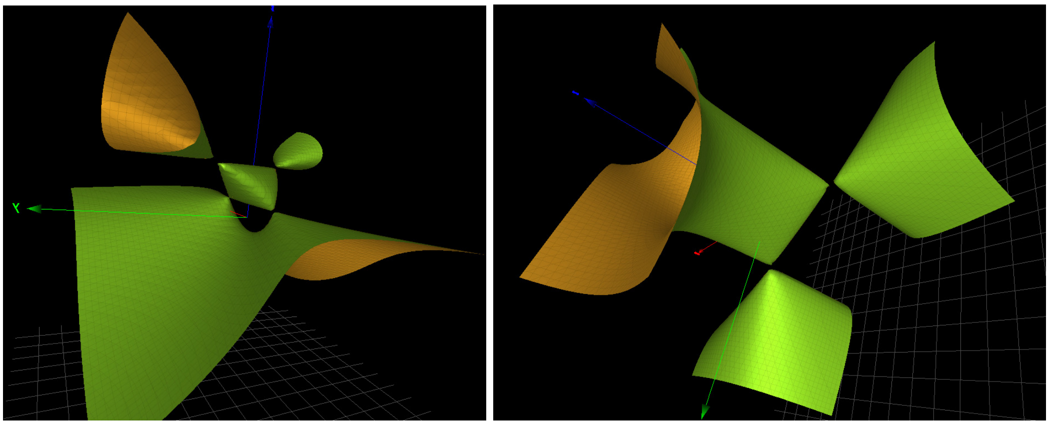

Figure 4. Left: the cubic surface $f_{4,\{\}}^{(2A_1)}(x,y,z)$. Right: the cubic surface $f_b^{(3A_1)}(x,y,z)$.

Table 2 lists a few selected transcription factors, their card seq and the corresponding singular surfaces, if any As announced, in the selected transcription factors, there exists a correlation between the lack of 'syntactical freedom', or the presence of a singular surface in the character variety, with an identified disease.

3.3. *Algebraic morphology of microRNAs*

MicroRNAs (miRNAs) play a fundamental role in the expression and regulation of genes by targeting specific messenger RNAs (mRNAs) for degradation or translational repression. The miRNAs are approximately 22 nt long single-stranded RNA molecules. The genes encoding miRNAs are much longer than the processed mature miRNA molecule. Many miRNAs are known to reside in introns of their pre-mRNA host genes and share their regulatory elements, primary transcript, and have a similar expression profile. MicroRNAs are transcribed by RNA polymerase II as large RNA precursors called pri-miRNAs. The pre-miRNAs are aproximately 70-nucleotides in length and are folded into imperfect stem-loop structures, see Figure 1 (Right) for an example.

Each miRNA is synthesized as a miRNA duplex comprising two strands (-5p and -3p). However, only one of the two strands becomes active and is selectively incorporated into the RNA-induced silencing complex in a process known as miRNA strand selection [31,32]. For details about the mirRNA sequences we use the Mir database [33,34].

Plant miRNAs usually have near-perfect pairing with their mRNA targets so that gene repression proceeds through cleavage of the target transcripts. In contrast, animal miRNAs are able to recognize their target mRNAs by using as few as 6 to 8 nucleotides (the seed region) which is not enough pairing for leading to cleavage of the target mRNAs. A given miRNA may have hundreds of different mRNA targets, and a given target might be regulated by multiple miRNAs.

Disregulation of miRNAs may lead to a disease like cancer. A key microRNA known as an oncommir (involved in immunity and cancer) is mir-155.

Specifically the -3p strand is mir-155-3p. Figure 5 (top) illustrates the complementary base-pairing between miR-155-3p and the human IRAK3 (interleukin-1 receptor-associated kinase 3) mRNA [10, Figure 4] and the relevant seed sequence UCCUAC. The card seq for this sequence is the two-letter free froup F_2 and the Groebner basis for the corresponding character variety contains the surface $f_b^{(A_1)}(x,y,z) = x^2 + y^2 - 6z^2 + 4xyz$ that has a single simple singularity as shown at the bottom of Figure 5. If one retains the full seed sequence is UCCUAC(A) then the card seq passes to that of the free group F_2 to the group π_2 and the singular surface is lost. This is a case where the 'bandwidth' of the seed is critical in the (dis)regulation of the miRNA. These results are transcribed in Table 3.

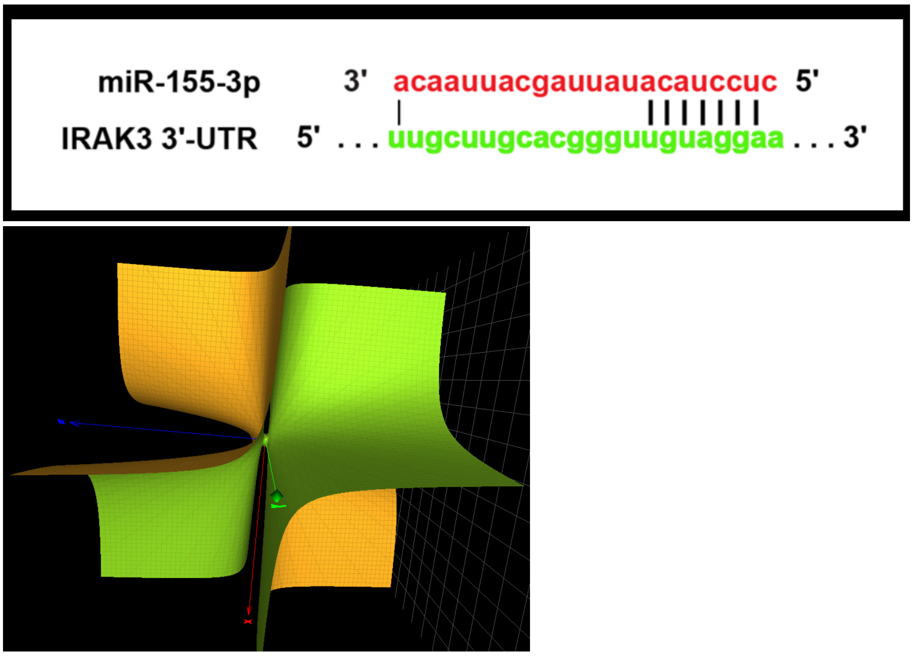
miR-155-3p
3' acaauuacgauuauacauccuc 5'
IRAK3 3'-UTR
5' . . . uugcuugcacggguuguaggaa . . . 3'

Figure 5. Up: Complementary base-pairing between miR-155-3p and the human Irak3 (interleukin-1 receptor-associated kinase 3) mRNA [10, Figure 5]. The requisite'seed sequence' base-pairing is denoted by the bold dashes. Down: the surface $f_b^{(A_1)}(x,y,z) = x^2 + y^2 - 6z^2 + 4xyz$.

For the case of -5p strand mir-155-3p, the seed sequence UUAAUGCUA contains four distinct letters. This case is similar to generic Groebner bases obtained from four letter seeds. Depending on the choices of parameters a,b,c,d,e,f,g,h, the Groebner basis contains the Cayley cubic $f_H^{(4)}(x,y,z)$, the Fricke surface $\kappa_3(x,y,z)$ (that is related to the link L6a1 [21, Figure 2]), the surface $f_a^{(3A_1)}(x,y,z)$ shown in Figure 2 and other surfaces. In this generic case, the surface is found with (at most) 4 copies where each copy is attached to a distinct puncture of the 4-punctured 4-sphere $S_{4,2}$.

In table 3, this generic case is denoted 4 × generic (or 3 × generic for mir-133-5p). These results are transcribed in Table 4.

A small list of huma miRNAs is investigated in Tables 3 and 4 corresponding to 3-letter and 4-letter seeds. the prefix 'hsa' is for the human specie. Like for transcription factors

Table 3. A few human (prefix 'hsa') microRNAs whose group structure is away from a free group or whose Groebner basis of the $SL(2,\mathbb{C})$ character variety contains a singular surface. The symbol mir is for the identification in the Mir database [34], seed is for the seed of the miRNA, card seq is for the cardinality sequence of conjugacy classes of subgroups of the group whose seed is the generator, sing is the identification of a singular surface within the Groebner basis, the last columnn is for a reference paper and the corresponding disease [31]. The card seq for π_1 and π_1' are given in [3, Table5]. The card seq for π_2' is $[1,3,7,34,139,931,5208,43867\cdots]$. For hsa-mir-124-1-3p, one encounters the Fricke surface $f_{2,\{\}}^{(A_1)} = xyz + x^2 + y^2 + z^2 - 2y$ in the character variety.

mir	seed	card seq	sing	ref & disease
hsa-mir-193b-5p	GGGGUU	π_1	no sing	[31,34]
	GGGGUUU	π_1'	no sing	lung cancer
hsa-mir-155-3p	UCCUAC	F_2	$f_b^{(A_1)}(x,y,z)$	[31,32,34]
	UCCUACA	π_2	no sing	multiple sclerosis
hsa-mir-193a-5p	GGGUCUU	F_2	$f_b^{(A_1)}(x,y,z)$	[31,34]
				breast cancer
hsa-mir-223-5p	GUGUAUU	.	.	.
hsa-mir-133-3p	UUGGUC	F_2	$f_b^{(3A_1)}(x,y,z)$	[31,34]
	UUGGUCC	π_2'	no sing	atrial fibrillation
hsa-mir-124-3p	AAGGCA	F_2	$f_b^{(3A_1)}, f_{2,\{\}}^{(A_1)}$	[34,35]
	AAGGCAC	.	no sing	Alzheimer's disease

Table 4. The opposite strand of the microRNA considered in Table 3. The seed sequence is made of 4 distinct bases and the corresponding card seq is the free group F_3 of rank 3. The Groebner basis contains 4 copies of the generic collection of surfaces $\kappa_4(x,y,z)$, $f^{(3A_1)}(x,y,z)$, $\kappa_3(x,y,z)$, etc, as shown in Figure 6, except for the -5p strand of mir-133 where there are only 3 copies of the generic surfaces.

mir	seed	card seq	sing	ref & disease
hsa-mir-193b-3p	ACUGGCC	F_3	4× generic	[31,34]
hsa-mir-155-5p	UUAAUGCUA	.	.	[31,32,34]
hsa-mir-193a-3p	ACUGGCC	.	.	[31,34]
hsa-mir-223-3p	GUCAGUU	.	.	.
hsa-mir-124-5p	GUGUUCA	.	.	.
hsa-mir-133-5p	GCUGGUA	.	3× generic	. [34,35]

in Table 2, the lack of 'syntactical freedom', or the occurrence of a singular surface in the character variety, is symptomatic of a disease.

4. Conclusion

We found in this work that a signature of a disease may be given in terms of the group structure of a DNA/RNA sequence and the related character variety representing the group. The DNA motif of a transcription factor, or the seed of a microRNA, defines the generator of a group π. As soon as π is away from a free group F_r (with $r+1$ the number of distinct bases in the sequence) or the $SL(2,\mathbb{C})$ character variety $\mathcal{G}$ of π contains singular surfaces with isolated singularities, a potential disease is on sight. One would like to be more predictive in identifying the potential disease with peculiar groups or singular surfaces.

First of all, most of the time, the surfaces encountered in the context of TFs and miRNAs are degree 3 del Pezzo, in contrast to surfaces obtained from other DNA sequences, as in Equation 4. But the degree 3 del Pezzo family is very rich. For instance, the singular surface $f_b^{(A_1)} = x^2 + y^2 - 6z^2 + 4xyz$ (see Figure 5) is part of the character variety of TF

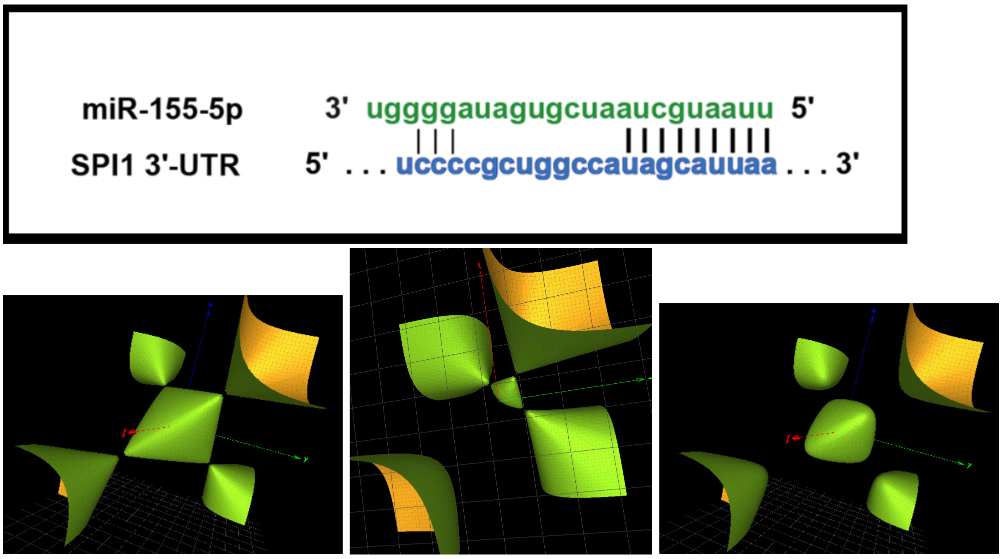
miR-155-5p
3' uggggauagugcuaaucguaauu 5'
SPI1 3'-UTR
5' . . . ucccgcuggccauagcauuaa . . . 3'

Figure 6. Up: Complementary base-pairing between miR-155-5p and the human Spi1 (spleen focus forming virus proviral integration oncogene) [10, Figure 4]. The requisite 'seed sequence' base-pairing is denoted by the bold dashes. Down (from left to right): the surfaces $f_H^{(4)} = \kappa_4(x,y,z)$, $f^{(3A_1)}(x,y,z)$ and $\kappa_3(x,y,z)$, four copies of them are contained within the Groebner basis for the character variety.

Pitx1 (see Table 2) and of miRNAs 155-3p and 193a-5p (in Table 3). Then, the singular surface $f_{2,\{\}}^{(A_1)} = x^2 + y^2 + z^2 + xyz - 2y$ is part of the character variety of mirRNA 133-3p. Both surfaces have a simple singular point of type A_1 but distinct singular subschemes (see Section 2 for the notation).

An exception to the degree 3 del Pezzo rule was found in investigating the character variety for the Prdm1 transcription factor in subsection 3.1.

Do these features and other ones to be described later may help for the diagnostic of a potential disease? There is room for much work in the future along these lines.

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

Funding: "This research received no external funding"

Informed Consent Statement: "Not applicable"

Data Availability Statement: "Data are available from the authors after a reasonable demand"

Conflicts of Interest: "The authors declare no conflict of interest"

MDPI Multidisciplinary Digital Publishing Institute

References

1. Planat, M.; Aschheim, R.; Amaral, M.M.; Fang, F.; Irwin, K. Graph coverings for investigating non local structures in protein, music and poems. *Science* **2021**, *3*, 39.
2. Irwin, K. The code-theoretic axiom; the third ontology. *Rep. Adv. Phys. Sci.* **2019**, *3*, 39.
3. 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.
4. Dang, Y.; Gao, J.; Wang, J.; Heffernan, R.; Hanson, J.; Paliwal, K.; Zhou, Y. Sixty-five years of the long march in protein secondary structure prediction: The final strech? *Brief. Bioinform.* **2018**, *19*, 482–494.
5. Bansal, A.; Kaushik, S.; Kukreti, S. Non-canonical DNA structures: Diversity and disease association. *Front. Genet.* **2022**, *13*, 959258. https://doi.org/10.3389/fgene.2022.959258.

6. Planat, M.; Aschheim, R.; Amaral, M. M.; Fang F.; Chester, D.; Irwin K. Character varieties and algebraic surfaces for the topology of quantum computing. *Symmetry* **2022**, 14, 915.
7. Planat, M.; Amaral, M.M.; Fang F.; Chester, D.; Aschheim, R.; Irwin, K. DNA sequence and structure under the prism of group theory and algebraic surfaces. *Int. J. Mol. Sci.* **2022**, *23*, 13290.
8. Gröbner basis. Available online: https://en.wikipedia.org/wiki/Gröbner_basis (accessed on 1 August 2022).
9. Martinez, N. J.; Walhout, A. J.M. The interplay between transcription factors and microRNAs in genome-scale regulatory networks. *Bioessays* **2009**, *31*, 435–445.
10. miR-155. Available online: https://en.wikipedia.org/wiki/MiR-155 (accessed on 18 November 2022).
11. Kwak, J.H.; Nedela, R. 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 November 2022).
13. Hehl, F.W.; von der Heyde,P.; Kerlick, G.D.; Nester, J.M., General relativity with spin and torsion: Foundations and prospects. *Rev. Mod. Phys.* **1976**, *48* (3), 393-416.
14. Rovelli, C., Vidotto, F.: Covariant Loop Quantum Gravity. Cambridge University Press 1 edition, (2014).
15. 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.
16. Ashley, C.; Burelle J. P.; Lawton, S. Rank 1 character varieties of finitely presented groups, *Geom. Dedicata* **2018**, *192*, 1–19.
17. Python code to compute character varieties. Available online: http://math.gmu.edu/s̃lawton3/Main.sagews (accessed on 1 May 2021).
18. Bosma, W.; Cannon, J. J.; Fieker, C. ; Steel, A. (eds). *Handbook of Magma functions*, Edition 2.23; University of Sydney: Sydney, Australia, 2017; 5914pp (accessed on 1 January 2019).
19. Cantat, S. Bers and Hénon, Painlevé and Schrödinger. **2009**, *Duke Math. J. 149*, 411–460.
20. Farb, B.; Margalit, D. *A primer on mapping class groups*; Princeton University Press: Princeton, New Jersey, USA, 2012.
21. Planat, M.; Chester, D.; Amaral, M.; Irwin K. Fricke topological qubits. *Quant. Rep.* **2022**, *4*, 523-532.
22. Benedetto, R. L.; Goldman W. M. The topology of the relative character varieties of a quadruply-punctured sphere. *Experiment. Math.* **1999**, *8*, 85–103.
23. Iwasaki, K. An area-preserving action of the modular group on cubic surfaces and the Painlevé VI. *Comm. Math. Phys.* **2003**, *242*, 185-219.
24. Inaba, M.; Iwasaki, K.; Saito, M. H. Dynamics of the sixth Painlevé equation. *arXiv* **2005**, arXiv:math.AG/0501007
25. Enriques Kodaira classification, available online: https://en.wikipedia.org/wiki/Enriques-Kodaira_classification
26. Doody, G. M.; Care, M. A.; Burgoyne, N. J.; Bradford, J. R.; Bota, M.; Bonifer, C.; Westhead, D. R.. An extended set of PRDM1/BLIMP1 target genes links binding motif type to dynamic repression ooze, R. M. *Nucl. Acids Res.* **2010**, *38* 5336–5350.
27. 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/ (accessed on 1 November 2022).
28. Jauch, R. Crystal tructure and DNA inding of the homeodomain of the stem cell transcription factor Nanog. *J. Mol. Biol.* **2008**, *376*, 758–770.
29. Schuff, M; Siegel D.; Philipp, M. Bunsschu, K.; Heymann, N.; Donow, C.; Knöchel, W. Characterization of Danio rerio Nanog and Functional Comparison to Xenopus Vents. *Stem cells and Devt.* **2012**, *21*, 1225–1238.
30. Schaeffer, L.N.; Huchet-Dymanus, M.; Changeux J.P. Implication of a multisubunit Ets-related transcription factor in synaptic expression of the nicotinic acetylcholine receptor. *EMBO J.* **1998**, *17*, 3078–3090.
31. Medley, C. M.; Panzade G.; Zinovyeva, A. Y. MicroRNA stran selection,: unwinding the rules. *WIREs RNA* **2021**, *12*, e1627.
32. Dawson, O.; Piccinini, A. M. miR-155-3p: processing by-product or rising star in immunity and cancer? *Open Biol.* **2022**, *12*, 220070.
33. Kozomara, A.; Birgaonu, M.; Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucl. Acids Res.* **2019**, *47*, D155–D162.
34. miRBase: the microRNA database. Available online: https://www.mirbase.org/ (accessed on 1 November 2022).
35. Kou, X.; Chen, D.; Chen, N. The regulation of microRNAs in Alzheimer's disease. *Front. Neurol.* **2020**, *11*, 288.