Group Theory of Messenger RNA Metabolism and Disease

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

[1]   Université de Franche-Comté, Institut FEMTO-ST CNRS UMR 6174, 15 B Avenue des Montboucons, 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: Our recent work has focused on the application of infinite group theory and related algebraic geometric tools in the context of transcription factors and microRNAs. We were able to differentiate between "healthy" nucleotide sequences and disrupted sequences that may be associated with various diseases. In this paper, we extend our efforts to the study of messenger RNA metabolism, showcasing the power of our approach. We investigate (i) mRNA translation in prokaryotes and eukaryotes, (ii) polyadenylation in eukaryotes, which is crucial for nuclear export, translation, stability, and splicing of mRNA, (iii) miRNAs involved in RNA silencing and post-transcriptional regulation of gene expression, and (iv) the identification of disrupted sequences that could lead to potential illnesses. To achieve this, we employ (a) infinite (finitely generated) groups f_p, with generators representing the $r + 1$ distinct nucleotides and a relation between them [e.g., the consensus sequence in the mRNA translation (i), the poly(A) tail in item (ii), and the miRNA seed in item (iii)], (b) aperiodicity theory, which connects "healthy groups" f_p to free groups F_r of rank r and their profinite completion $\hat{F}_r$, and (c) the representation theory of groups f_p over the space-time-spin group $SL_2(\mathbb{C})$, highlighting the role of surfaces with isolated singularities in the character variety. Our approach could potentially contribute to the understanding of the molecular mechanisms underlying various diseases and help develop new diagnostic or therapeutic strategies.

Keywords: RNA metabolism; micro-RNAs; diseases; finitely generated group; $SL(2, \mathbb{C})$ character variety; aperiodicity

---

1. Introduction

Genome-scale metabolic pathways [1,2], genome–environment interactions [3], the immune response [4], post-transcriptional regulatory mechanisms [5] and oncohistones [6] represent aspects of a research field connecting the heritable genetic code to other biological codes.

The aforementioned genetic code is defined precisely as a non-injective map from the 64 codons to the 20 amino acids. Both finite groups [7–9] and quantum groups [10] play a leading role in modeling this code.

We refer to the "epigenetic code" as all processes that reveal and execute gene expression. This includes DNA methylation processes [11], mRNA translation preparation, the poly(A) tail, the RNA-induced silencing complex (RISC) — a vital tool in gene regulation comprising single strands of RNA (ssRNA) and double strands of small interfering RNA (siRNA) — and other regulatory nucleotide sequence fragments that are discarded after splicing. For a relation between the epigenetic code and morphogenesis, see [12].

For studying the epigenetic code (hereinafter referred to as the e-code), we employ infinite (finitely generated) groups denoted by f_p, and their representations over the (2×2) matrix group $SL_2(\mathbb{C})$,

where the entries are complex numbers [13]. The significance of this group extends across all fields of physics as it represents a space-time-spin group. In this study, we apply a mathematical field known as algebraic geometry to define the e-code.

Our crucial observation is that an f_p group associated with a "healthy" sequence usually approximates a free group F_r, where the rank r equals the number of distinct nucleotides (nt) minus one. A sequence deviating from this may suggest a potential e-code deregulation leading to a disease. However, an f_p group closely resembling a free group does not provide sufficient assurance against a disease. Additional examination of the $SL_2(\mathbb{C})$ representations of f_p — termed the character variety— specifically, its basis — called a Groebner basis $\mathcal{G}$ — is necessary.

The Groebner basis comprises a set of surfaces. A surface within $\mathcal{G}$ containing isolated singularities indicates another potential disease that can be identified specifically, e.g., relating to an oncogene or a neurological disorder [13, Figure 6, Tables 2 to 4]. The e-code we define comprises such algebraic geometric characteristics.

An additional attribute of "healthy" sequences, which leads to a group f_p approximating the free group F_r and not mentioned in [13], is their connection to aperiodicity. Schrödinger's book [14] proposes aperiodicity of living "crystals". Our paper [15] characterizes aperiodic DNA sequences. We further this concept by introducing the so-called profinite completion $\hat{F}_r$ of the free group F_r. A sequence $f_p^{(l)}$ of finitely generated groups approaching F_r emerges by applying l repeated substitutions to the generators of f_p. However, all distinct groups $f_p^{(l)}$ should possess the same profinite completion $\hat{F}_r$. Profinite groups $\hat{F}_1$ (corresponding to sequences containing two distinct nt) and $\hat{F}_2$ (corresponding to sequences containing three distinct nt) have been examined in a prominent algebraic geometry treatise [16]. We present the details below in a manner that is accessible to a non-specialist reader.

In Section 2, we illustrate our mathematical concepts through a few simple pedagogical examples. In Section 3, we apply these concepts to the cases of mRNA translation and microRNAs. In Section 4, we provide additional comments, a summary diagram and perspectives.

2. Methods and preliminary results

2.1. Infinite finitely generated groups f_p and free groups F_r

The TATA box

We'll start with a simple example of an infinitely finitely generated group taken from the context of introns. The DNA sequence located in the core promoter region of many eukaryotic genes is the Goldberg-Hogness sequence, also known as the TATA box. This sequence contains a non-coding segment with repeated T and A base pairs. The TATA box serves as the binding site for the TATA-binding protein and other transcription factors in some eukaryotic genes. Its consensus sequence takes the form rel=TATAAAA. Variations in this consensus sequence, resulting from genetic polymorphism, can lead to diseases like Gilbert's syndrome and immune suppression [17].

In our methodology, we define the group $f_p = \langle A, T | \text{rel} \rangle$, which contains infinitely many elements. There are numerous ways to investigate this group, but we opted for a specific one. This method involves calculating the number of conjugacy classes of subgroups of index d of f_p (a sequence we'll refer to as the card seq of f_p). The card seq of f_p for the selected TATA sequence is $[1, 1, 2, 3, 2, 8, 7, 10, 18, 28 \cdots]$. Interestingly, the group $H_3 = \langle A, T | A^2 = T^3 \rangle$ shows a similar card seq (at least up to the highest index we can reach with the calculations). The group H_3, as defined, is isomorphic to the so-called modular group $PSL(2, \mathbb{Z})$ – the group of (2×2) matrices of determinant 1 with integer entries. This group has an intriguing topological interpretation as the fundamental group of the trefoil knot manifold. Thus, we find that the group f_p is 'close' to H_3 since the card seq of both groups is the same, but we can easily verify that f_p and H_3 are not isomorphic.

In paper [19, Section 3.1 and Table 2], we discovered that Hecke groups $H_q = \langle A, T | A^2 = T^q \rangle$, with $q = 3$ or 4, have a card seq corresponding to 'healthy' TATA box sequences. The f_p group for a TATA box with a card seq resembling that of Hecke groups, with $q \neq 3$ or $q \neq 4$, or even that of groups slightly different from H_3 and H_4, signifies Gilbert's syndrome.

Polyadenylation signals

For our second example, we select a sequence from the context of eukaryotic polyadenylation [18]. Polyadenylation involves the addition of a poly(A) tail to an RNA transcript, usually a messenger RNA (mRNA). A consensus poly(A) sequence takes the form rel1=AAUAAA. This corresponds to a two-generator group of the form $f_p = \langle AU | \text{rel1} \rangle$. The card seq of such a group is found to be $[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, \cdots]$, implying a single conjugacy class for each index. It appears that the free group $F_1 = \langle A, U | AU \rangle$, of rank 1, has the same card seq as the f_p group with relation rel1, even though both groups are not isomorphic.

Another consensus poly(A) sequence takes the form rel2=UGUAA. This corresponds to a three-generator group of the form $f_p = \langle A, U, G | \text{rel2} \rangle$. The card seq of such a group is found to be $[1, 3, 7, 26, 97, 624, 4163, \cdots]$. Intriguingly, the free group $F_2 = \langle A, U, G | AUG \rangle$, of rank 2, has the same card seq as the f_p group with relation rel2, despite both groups not being isomorphic.

From our perspective, DNA/RNA sequences that lead to f_p groups closely resembling a free group are considered 'healthy' sequences [13,15,19]. The standard poly(A) sequences mentioned earlier play a regulatory role in producing mature mRNA during translation. Sequences that generate an f_p group diverging from a free group F_r may be indicative of a disease.

2.2. Aperiodic sequences, their attached groups f_p and free groups

In this subsection, we'll elucidate how a group f_p, with a card seq identified to be close to a free group F_r, can be linked to an aperiodic sequence and the profinite completion $\hat{F}_r$. We introduced the concept of aperiodic groups and sequences in our earlier papers [19, Section 4] and [15, Section 2].

Consider the motif rel $= TTTATTA$, which serves as a consensus sequence for the transcription factor of the DBX gene in Drosophila melanogaster (fruit flies). This gene is involved in neuronal specification and differentiation. The group $f_p = \langle A, T | \text{rel} \rangle$ has the same card seq as the free group F_1 of rank 1. Furthermore, by splitting rel into two segments rel $= \text{rel}_A \text{rel}_T$ and applying the substitution maps $A \to \text{rel}_A = TTTA$, $T \to \text{rel}_T = TTA$, we generate the substitution sequence

$S_{DBX} = A, T, AT, TTTATTA, TTATTATTATTTATTATTATTTA, \cdots$. Upon inspection, it's straightforward to observe that all finitely generated groups $f_p^{(l)}$, with their generators being

$AT, TTTATTA, TTATTATTATTTATTATTATTTA, \cdots$, respectively, possess the card seq of F_1.

As per Reference [19, Section 4], a substitution rule to be considered aperiodic must satisfy two conditions: (1) the substitution matrix M must be primitive, meaning it should be a strictly positive matrix (all entries > 0), irreducible, and M^k should be strictly positive for some k. This condition is denoted as $M >> 0$, and (2) the Perron-Frobenius λ_{PF} eigenvalue must be irrational. It's worth noting that the Perron-Frobenius eigenvector of an irreducible non-negative matrix is the only one whose entries are all positive.

The aforementioned sequence has a substitution matrix $M = \begin{pmatrix} 1 & 3 \\ 1 & 2 \end{pmatrix}$. One can verify that M is primitive since $M^2 >> 0$ and $\lambda_{PF} = (3 + \sqrt{13})/2$. Conditions (1) and (2) are satisfied, implying that the substitution S_{DBX} is aperiodic.

Numerous other genes have transcription factors with a motif rel generating an aperiodic sequence [15, Table 2].

2.3. Aperiodic sequences and the profinite groups $\hat{F}_r$

This section can be skipped without affecting the comprehension of the rest of the paper. It endeavors to answer the following question: why do the aforementioned groups $f_p^{(l)}$ produce the same card seq as that of the free group F_1? The tentative answer to this question is that the profinite completion of all groups $f_p^{(l)}$ is the profinite group $\hat{F}_1$. By making this observation, we align the aperiodicity of sequences with profinite groups. Profinite groups were introduced by Grothendieck in the context of algebraic geometry [16]. Here, we describe the necessary ingredients for the layperson, focusing first on $\hat{F}_1$ and then on $\hat{F}_2$, and their relevance to our present work.

A group G can be considered a 'topological group' by applying the 'discrete topology', in which the elements of G are points of a 'discrete space', form a 'discontinuous sequence' and are isolated from each other. Every subset is 'open' in the discrete topology. A profinite group is a topological group that, in a certain sense, is assembled from a system of finite groups. A profinite group requires a system of finite groups and group homomorphisms between them.

Given a group G, there is a related profinite group $\hat{G}$ defined as the inverse limit $\hat{G} = \lim_{\leftarrow} G/N$, of the groups G/N, where N runs through the normal subgroups of G of finite index [a normal subgroup is a subgroup that remains invariant under conjugation by members of the group]. Each finite quotient group corresponds to a normal subgroup N of G and the profinite completion $\hat{G}$ can be perceived as containing an analogue of each of these normal subgroups.

The profinite group $\hat{G}$ exhibits several properties: it is non-abelian, residually finite [meaning that for any non-identity element g in $\hat{G}$, there exists a finite quotient of $\hat{G}$ in which g is not the identity], and totally disconnected [meaning that the only connected subsets of $\hat{G}$ are singletons, sets containing only one element].

In general, an explicit construction of profinite groups $\hat{G}$ cannot be obtained. However, $\hat{F}_1$ and $\hat{F}_2$ are not overly complex to handle.

About the profinite group $\hat{F}_1$

Let's begin with $\hat{F}_1$. The free group F_1 on a single generator can be described as a group with one generator, say a, and no relations. It consists of all possible finite strings that can be formed by combining the generator and its inverse. It is the infinite cyclic group $Z = \{1, a, a^{-1}, a^2, a^{-2}, a^3, a^{-3}, \cdots\}$. Now, let's discuss the profinite completion of F_1. The profinite group $\hat{F}_1$ is isomorphic to the group of all units of the commutative ring of p-adic integers Z_p, across all primes p. It is often denoted as Z_p^* since it corresponds to the elements of Z_p with a valuation of zero. The p-adic integers are a special class of numbers utilized in number theory and algebraic geometry.

About the profinite group $\hat{F}_2$

Let's briefly discuss $\hat{F}_2$. This topic was initiated in [16]. The subject is deep and complex. It's connected to the so-called Belyi theorem, a fundamental result that establishes a connection between algebraic curves defined over the algebraic closure of the rationals, $\bar{\mathbb{Q}}$, and certain rational functions called Belyi functions.

An algebraic curve defined over $\bar{\mathbb{Q}}$ can be represented as a branched covering of the Riemann sphere (the complex projective line $\mathbb{P}^1(\mathbb{C})$) branched only over three points (usually taken as 0, 1, and ∞) if and only if the curve itself is defined over a number field, which is a finite extension of the field of rational numbers $\mathbb{Q}$.

In other words, the Belyi theorem implies that an algebraic curve defined over a number field can be mapped to the Riemann sphere in such a way that the ramification (branching) is restricted to just three points. The rational functions that provide these branched coverings are known as Belyi functions.

The significance of the Belyi theorem lies in the fact that it provides a method to study algebraic curves defined over number fields by analyzing their ramified coverings and the associated dessins d'enfants, which are combinatorial objects encoding the ramification data.

Specifically, we have the crucial result that

$$\hat{\pi}_1(\mathbb{P}^1(\mathbb{C}) \setminus \{0, 1, \infty\}) \cong \hat{F}_2,$$

i.e., the so-called étale fundamental group for the triply branched projective line is the profinite group $\hat{F}_2$.

2.4. $SL_2(\mathbb{C})$ representations of groups f_p and a Groebner basis $\mathcal{G}$

While the previous section about profinite groups showcases the importance of algebraic geometry in the context of DNA/RNA sequences, it remains somewhat abstract. To address this, we can consider the representations of an f_p group over the space-time-spin group $SL_2(\mathbb{C})$, as we did in [13,15].

Representations of f_p in $SL_2(\mathbb{C})$ are homomorphisms $\rho : f_p \to SL_2(\mathbb{C})$ with character $\kappa_\rho(g) = \text{tr}(\rho(g))$, $g \in f_p$. The notation $\text{tr}(\rho(g))$ signifies the trace of the matrix $\rho(g)$. The set of characters is employed to determine 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[20,21].

In our paper [13, Section 2.2], we elaborated that the character variety of f_p is a set comprised of a sequence X of multivariate polynomials. A particular basis related to X is the Groebner basis $\mathcal{G}(X)$, whose factors define hypersurfaces.

For a two-generator group f_p, the factors are three-dimensional surfaces. In general, these surfaces can be classified by mapping them to a rational surface across five categories [13, Section 3]. Often encountered surfaces are degree p Del Pezzo surfaces where $1 \le p \le 9$. A rational surface may either be non-singular, 'almost non-singular', having only isolated singularities, or singular. Almost non-singular surfaces are crucial in our context. A simple singularity is referred to as an A-D-E singularity and must be of the type A_n, $n \ge 1$, D_n, $n \ge 4$, E_6, E_7 or E_8.

The A-D-E type is mirrored in the notation we employ. For instance, $S^{(lA_1, mA_2, nA_3, \cdots)}$ denotes a surface containing l type A_1, m type A_2, n type A_3 singularities, etc. A generic surface is the Cayley cubic we encountered in our previous papers, defined as $S^{(4A_1)} = xyz + x^2 + y^2 + z^2 - 4$ [13, Figure 5].

For a three-generator group f_p, the factors of $\mathcal{G}(X)$ are seven-dimensional surfaces of the form $S_{a,b,c,d}(x, y, z)$. Some of them belong to the Fricke family [13, Equation 3], which is associated with the four-punctured sphere. But for a chosen set of parameters a, b, c, d, the hypersurface reduces to an ordinary three-dimensional surface.

For a four-generator group f_p, the factors of $\mathcal{G}(X)$ are 14-dimensional surfaces containing 4 copies of the form $S(x, y, z)$, $S(x, u, v)$, $S(y, u, v)$ and $S(z, v, w)$ for selected choices of 8 parameters.

Groebner basis for the TATA box

The Groebner basis for the character variety associated with the f_p group of generator rel=TATAAAA of the TATA box, studied in subsection 2.1, is found to be:

$$\mathcal{G}_{TATA} = (z^4 - xy^2 - xyz + x^2 + y^2 + yz - 3z^2 + x - 2)(x^2z - xy - xz + y - z)$$
$$S^{(A_2)} S^{(A_4)} (x^3 - z^2 - 3x + 2),$$

where $S^{(A_2)} = x^2y - z^3 - xz - y + 3z$ and $S^{(A_4)} = xz^2 - x^2 - yz - x + 2$ are degree 3 Del Pezzo surfaces.

The Groebner basis $\mathcal{G}_{TATA}$ comprises a degree 2 Del Pezzo surface (see Figure 1, up), and a rational scroll whole analytic expression is in the first row. Both surfaces are singular. The second row consists

of two surfaces with simple singularities of type A_2 and A_4, respectively. The last term represents a curve (not a surface).
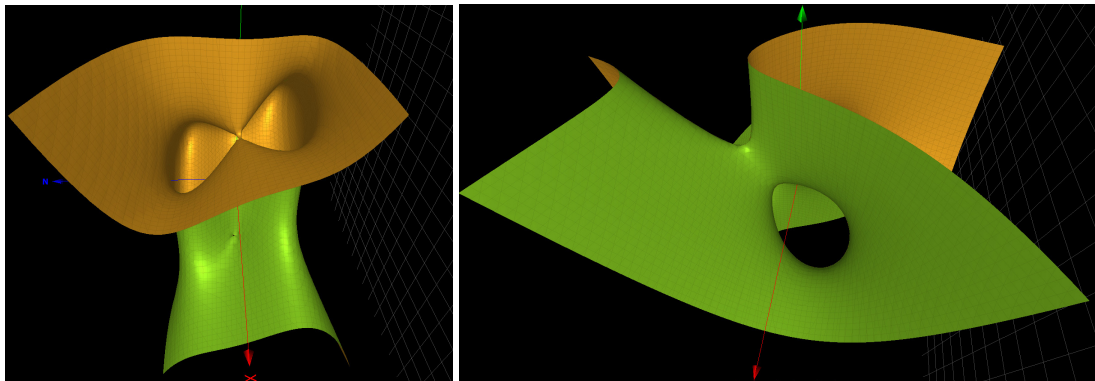
Figure 1. Up: The degree 2 Del Pezzo surface within $\mathcal{G}_{TATA}$. Down: The degree 3 Del Pezzo surface $S^{(A_1)}$ within $\mathcal{G}_{rel1}$.

Groebner basis for polyadenylation signals

For the first polyadenylation signal considered in subsection 2.1, the relation of the f_p group is rel1=AAUAAA. The corresponding Groebner basis is:

$$\mathcal{G}_{rel1} = 3 \text{ rational scrolls} \times P^2 \times S^{(4A_1)} S^{(A_1)} \times \text{curve}.$$

The Groebner basis $\mathcal{G}_{rel1}$ contains three rational scrolls, a surface birationally equivalent to the projective plane P^2, the Cayley cubic $S^{(4A_1)}$, the degree 3 Del Pezzo surface $S^{(A_1)} = x^2y - xz^2 - xz + yz + x - y$ (see Figure 1, down) and a curve.

For the second polyadenylation signal considered in subsection 2.1, the relation of the f_p group is rel2=UGUAA. The factors of $\mathcal{G}(X)$ are seven-dimensional hypersurfaces $S_{a,b,c,d}(x, y, z)$. However, choosing specific parameters, such as $S_{0,0,0,0}(x, y, z)$ or $S_{1,1,1,1}(x, y, z)$, we obtain three-dimensional surfaces. These are found to be degree 3 Del Pezzo surfaces with simple singularities of the form S^{lA_2}, with l=1, 2 or 3, quadrics, or curves.

Groebner basis for the transcription factor of DBX gene

For the DBX gene studied in Section 2.2, the Groebner basis takes the form
$\mathcal{G}_{DBX} = \text{scroll} \times P^2 \times S^{(A_4)} \times S^{(A_2)} \times S^{(4A_1)} \times \text{curve}$, where scroll$=y^2z - xy - yz + x - z$ and $P^2 = z^4 - x^2y + xz - 4z^2 + y + 2$ are singular. The other factors are DP^3 surfaces with isolated singularities that are $S^{(A_4)} = yz^2 - y^2 - xz - y^2$, $S^{(A_2)} = z^3 - xy^2 + yz + x - 3z$, the Cayley cubic $S^{(4A_1)}$ and curve$=y^3 - z^2 - 3y + 2$.

3. Further results

In this section, we produce further results related to mRNA metabolism and miRNA.

3.1. Algebraic geometry of mRNA translation

The Shine-Dalgarno box

Ribosomal RNA (rRNA) – a type of non coding RNA– is the main component of a macromolecular machine, called the ribosome, whose role is to ensure mRNA translation. The initiation of translation needs the recognition of the appropriate sequences on the m-RNA by the ribosome. A major factor in this recognition is an mRNA-rRNA interaction first proposed by Shine and Dalgarno [22]. They proposed that the ribosomal nucleotides recognize the complementary purine-rich sequence

rel3=AGGAGGU, which is found around 8 bases upstream of the start codon AUG in a number of mRNAs found in viruses that affect Escherichia coli.

Let us study the group $f_p = \langle A, G, U | rel3 \rangle$. The card seq of f_p is found to be the same than that of the free group F_2.

The $SL_2(\mathbb{C})$ chararacter variety is a scheme X whose a Groebner basis $\mathcal{G}(X)$ is made of of 7-dimensional surfaces $S_{a,b,c,d}(x,y,z)$. By projecting to 3 dimensions, one gets surfaces like $S_{0,0,0,0}(x,y,z)$ and $S_{1,1,1,1}(x,y,z)$ as in Section 2.4. We find degree 3 Del Pezzo surfaces with isolated singularities $S^{(A_1)} = x^2y + yz^2 + 4xz + 4y$ and $x^2y + yz^2 + x^2 + z^2 + 6xz + 5y - 6z - 7$, $S^{(A_2)} = xyz + 2x^2 + z^2 + 4$ and $S^{(A_4)} = xyz + 3x^2 + z^2 - 5z$, quadrics and curves.

Kozak consensus sequence

The Kozak consensus sequence is a nucleotide motif that functions as the protein translation initiation site in most eukaryotic mRNA transcripts [23]. The small (40S) subunit of eukaryotic ribosomes bind, initially at the capped 5'-end of messenger RNA and then migrate, stopping at the first AUG codon in a favorable context for initiating translation. In eukaryotes, the Kozak sequence ensures that a protein is correctly translated from the genetic message, mediating ribosome assembly and translation initiation. A sequence logo of the most conserved bases around the initiation codon AUG for human mRNAs may be found in the first caption of [24] as rel4 $= ACCAUGGC$.

Let us study the group $f_p = \langle A, C, G, U | rel4 \rangle$. The card seq of f_p is found to be the same than that of the free group F_3 of rank 3. This group can be linked to an aperiodic sequence by following the steps given in Section 2.2.

By splitting rel4 into four segments rel4 $= rel_A rel_C rel_G rel_U$ and applying the substitution maps $C \to rel_C = A$, $A \to rel_A = CCAUG$, $U \to rel_U = G$, $G \to rel_G = C$, we generate the substitution sequence

$S_{Kozak} = C, A, U, G, CAUG, ACCAUGGC, CCAUGA^2CCAUGGC^2A, \cdots$.

Upon inspection, it's straightforward to observe that all finitely generated groups $f_p^{(I)}$, with their generators being

$CAUG, ACCAUGGC, CCAUGA^2CCAUGGC^2A, \cdots$, respectively, possess the card seq of F_3.

The aforementioned sequence has a substitution matrix $M = \begin{pmatrix} 0 & 2 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$. One can verify that

M is primitive since $M^4 >> 0$ and $\lambda_{PF} \approx 2.2055694$ is the only real (and irrational) solution of the equation $x^3 - 2x^2 - 1$. Conditions (1) and (2) of Section 2.2 are satisfied, implying that the substitution S_{Kozak} is aperiodic. See [25] for a connection of the later Perron-Frobenius eigenvalue to random Fibonacci sequences.

Mutation of a purine at position -3 with respect to the AUG codon is kwown to be associated to a disease such as a type of thalassemia due to a bad initiation of α-globin [23]. In our approach the mutation from rel4 to rel4'=CCCAUGGC leads to a substitution M' that is no longer primitive so that the property of aperiodicity of the sequence is lost. However the cardseq of the associated f_p group is still that of the free group F_3. No other substitution in the sequence rel4' can be found to restore the aperiodicity.

3.2. Algebraic geometry of miRNAs

A microRNA (miRNA) is a small, single-stranded, non-coding RNA molecule containing approximately 22 nucleotides. miRNAs play crucial roles in RNA silencing and post-transcriptional regulation of gene expression by specifically targeting certain mRNAs for degradation and translational repression [26,27]. miRNA genes are typically transcribed by RNA polymerase II (Pol II), which binds to a promoter located near the DNA sequence, encoding what will become the hairpin loop of a

pre-miRNA (for precursor-miRNA). The pre-miRNAs are approximately 70 nucleotides in length and fold into imperfect stem-loop structures.

Each miRNA is synthesized as a miRNA duplex comprising two strands (-5p and -3p). However, only one of the two strands is selectively incorporated into the RNA-induced silencing complex to act as a template for the transcript of a complementary mRNA [28,29]. For details about the miRNA sequences, we use the Mir database [30–32].

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

For previous results about how to define a f_p group from the seed of a miRNA, the reader may consult [13, Section 4.3].

Below, we focus on other examples.

miRNA hsa-mir-503

The slowest evolving miRNA gene in the human species (hsa) is hsa-mir-503 [31]. It regulates gene expression in various pathological processes of diseases, including carcinogenesis, angiogenesis, tissue fibrosis, and oxidative stress [33].

The seed region for mir-503-5p is seed1=AGCAGCGG. The corresponding group $f_p = \langle A, C, G | \text{seed1} \rangle$ has the card seq of the free group F_2. For this group, the Groebner basis with parameters $(a, b, c, d) = (0, 0, 0, 0)$ is quite simple: $\mathcal{G}_{mir-503-5p}(0, 0, 0, 0) = S^{(4A_1)}(x, y, z)$, which is the already mentioned Cayley cubic.

For $(a, b, c, d) = (1, 1, 0, 0)$, $\mathcal{G}_{mir-503-5p}(1, 1, 0, 0) = -3xyz\kappa_3(x, y, z)$, where $\kappa_3(x, y, z)$ is the Fricke surface found in [34, Section 3.3]. For $(a, b, c, d) = (1, 1, 1, 1)$, there are several more polynomials. One of them defines the Fricke surface $xyz + x^2 + y^2 + z^2 - 2x - y - 2$.

The considered seed region for mir-503-3p is GGGUAUU. The surfaces in the Groebner basis are very simple in this case, and no simple singularities exist within them.

miRNA hsa-mir-146a

Mir-146 is primarily involved in the regulation of inflammation and other processes functioning in the innate immune system. It plays a role in neuropathogenesis.

The considered seed region for hsa-mir-146a-5p is seed2=GAGAAC [31]. Again the corresponding group $f_p = \langle A, C, G | \text{seed2} \rangle$ has the card seq of the free group F_2.

The Groebner basis with parameters $(a, b, c, d) = (0, 0, 0, 0)$ is

$\mathcal{G}_{hsa-146a-5p}(1, 1, 1, 1) = (xz + y + 2)(y - z^2 + 2)^2(x^2 + z^2 - 2y - 4)S^{(3A_2)}$, where $S^{(3A_2)} = z^3 - xy - 2yz - 2x - 4z$.

The Groebner basis with parameters $(a, b, c, d) = (1, 1, 1, 1)$ is of the form

$\mathcal{G}_{hsa-146a-5p}(1, 1, 1, 1) = DP^4 \times f^{(2A_2)} \times$ quadric $\times$ curves, where DP^4 is a degree 4 del Pezzo surface.

miRNAs and disease

As we found earlier, a potential disease is associated with f_p groups whose character variety has a Groebner basis containing isolated singularities, even though the f_p group has the card seq of a free group [13, Figure 6]. This is the case for the latter two miRNAs. Additional examples can be found in [13, Table 3].

Besides isolated singularities, the Groebner basis may contain singular surfaces that are not simply singular. The DP^4 surface in $\mathcal{G}_{hsa-146a-5p}(1, 1, 1, 1)$ is an example of a singular surface. Further mathematical techniques are required to investigate these surfaces [35]. However, we will not discuss these methods in this paper.

4. Discussion

In this section, we summarize our paper by referring to the diagram in Figure 2. Given a short DNA/RNA sequence, rel, which represents a consensus sequence in a transcription factor, the seed of a miRNA, or a relevant sequence in mRNA recognition and processing, we construct a finitely generated group, f_p. The architecture of subgroups, card seq, within this group is computed (see Section 2.1). If the f_p card seq matches that of the free group F_r (of rank r equal to nt-1), we proceed to path 4; otherwise, a potential disease could be in sight (path 3). After reaching path 4, the next step involves checking the aperiodicity of rel and the corresponding f_p group, as described in Section 2.2. The final step is to examine the presence (or absence) of isolated singularities in the Groebner basis $\mathcal{G}$ for the $SL_2(\mathbb{C})$ character variety associated with f_p, as outlined in Section 2.4. For a healthy sequence, the path concludes at 6, while a potential disease may be indicated if the path ends at 3, 7 or 8.

In Table 1, we provide several examples of paths. All three checks can be performed, even if paths 4 or 5 are not followed. For instance, the termination {7,8} signifies that the sequence fails both in being aperiodic and in being devoid of simple singularities.

For sequences with 4 nt (like the sequence of transcription factor DBX or the Shine-Dalgarno sequence rel3), it is difficult to conclude about the risk of a disease. The generic Groebner basis $\mathcal{G}(x,y,z)$ always contains surfaces with isolated singularitis such as $S^{(4A_1)}$ and $S^{(3A_1)}$ and there are four copies of them. The termination {6,8} applies for these case.

Our approach is quite comprehensive and can be applied in numerous contexts beyond those we have considered thus far. It has the potential to impact the search for underlying causes of diseases and aid in the discovery of therapeutic strategies. The e-code, the processes that reveal and execute gene expression, has a sophisticated structure, which our mathematical approach aims to elucidate.
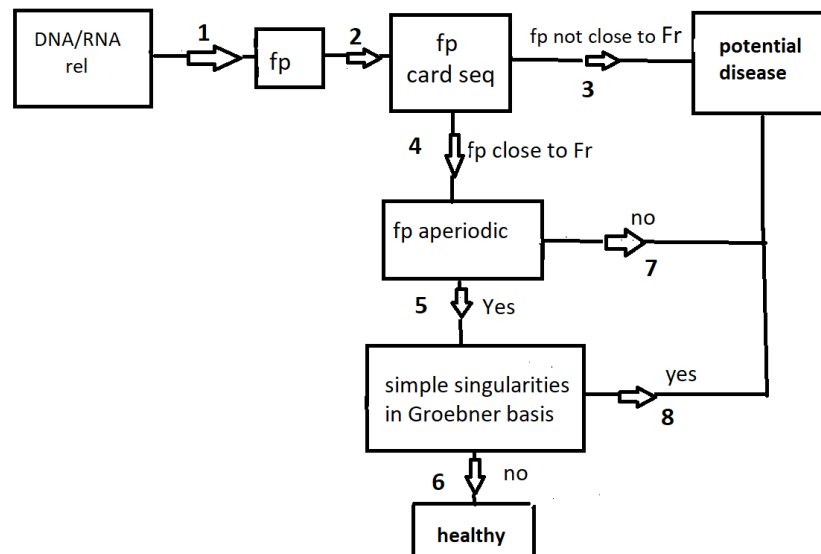
Figure 2. A diagram illustrating the main results discussed in the text. For example, for the transcription factor of the gene EGR1, rel=GCGTGGGCG [19, Section 4.1.2], the path is $1 \rightarrow 2 \rightarrow 4 \rightarrow 5 \rightarrow 6$ showing no risk of disease. But for the transcription factor of gene DBX (see Section 2.2 and 2.4), rel= TTTATTA, the path is $1 \rightarrow 2 \rightarrow 4 \rightarrow 5 \rightarrow 8$ meaning a potential disease (see Table 1).

Table 1. A few possible paths in the diagram of Figure 2 terminating at 6 (healthy) or (3)-(7)-(8) (potential disease). The selected examples are taken in three parts that are transcription factors (group 1), regulating elements in introns (group 2) and miRNAs (group 3). Details are given in the text. Otherwise a reference is provided.

Sequence	rel	path
EGR1 [19]	GCGTGGGCG	$1 \to 2 \to 4 \to 5 \to 6$
FOS [19]	TGAGTCA	$1 \to 2 \to 4 \to 5 \to \{6,8\}$
Nanog [19]	TAATGG	$1 \to 2 \to \{7,8\}$
DBX	TTTATTA	$1 \to 2 \to 4 \to 5 \to 8$
TATA	TATAAAA	$1 \to 2 \to 3 \to (7,8)$
poly(A) (rel1)	AAUAAA	$1 \to 2 \to 4 \to \{7,8\}$
poly(A) (rel2)	UGUAA	$1 \to 2 \to 4 \to \{7,8\}$
Shine-Dalgarno (rel3)	AGGAGGU	$1 \to 2 \to 4 \to 5 \to 8$
Kozak (rel4)	ACCAUGGC	$1 \to 2 \to 4 \to 5 \to \{6,8\}$
Kozak (rel4')	CCCAUGGC	$1 \to 2 \to 7$
hsa-mir-132-5p [36]	CCGUGGC	$1 \to 2 \to 4 \to 5 \to 6$
hsa-mir-503-5p (seed1) [33]	AGCAGCGG	$1 \to 2 \to 5 \to 8$
hsa-mir-146a-5p (seed2) [37]	GAGAAC	$1 \to 2 \to \{7,8\}$
hsa-mir-7-5p [38]	GGAAGA	$1 \to 2 \to \{3,7,8\}$
hsa-mir-7-5p	GGAAGAC	$1 \to 2 \to 4 \to 5 \to 6$
hsa-mir-7-3p	AACAAAU	$1 \to 2 \to 7$
hsa-mir-155-3p [29,37]	UCCUAC	$1 \to 2 \to 4 \to \{7,8\}$
hsa-mir-155-3p	UCCUACA	$1 \to 2 \to \{3,7\}$

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.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Computational data are available from the authors.

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

References

1.   Gu, C.; Kim G. B.; Kim, W. J.;Kim, H. U.; Lee, S. Y. Current status and applications of genome-scale metabolic models. *Genome Biology* **2019**, *20*, 121.

2.   Romão, L. mRNA metabolism in health and disease. *Biomedicines* **2022**, *10*, 2262.

3.   Peedicayil, J. Genome–environment interactions and psychiatric disorders. *Biomedicines* **2023**, *11*, 1209.

4.   Scharf, S.; Ackerman, J.; Bender, L.; Wurzel, P.; Schäfer, H.; Hansmann, M. L.; Koch, I. Holistic view on the structure of immune response: Petri net model. *Biomedicines* **2023**, *11*, 452.

5.   Marques, A. R.; Santos, J. X.; Martiniano, H.; Vilela, J.; Rasga, C.; Romão, L.; Vicente, A. M. Gene variants involved in nonsense-mediated mRNA decay suggest a role in autism spectrum disorder. *Biomedicines* **2022**, *10*, 665.

6.   Wan, Y. C.E. Histone H2B Mutations in Cancer. *Biomedicines* **2021**, *9*, 694.

7.   Fimmel, E.; Giannerini, S.; Gonzalez, D. L.; Strüngmann. Circular codes, symmetries and transformations. *J. Math. Biol.* **2014**, *70*, 1623–16434.

8.   Planat, M.; Aschheim, R; Amaral, M. M; Fang, F; Irwin, K. Complete quantum information in the DNA genetic code. *Symmetry* **2020**, *12*, 1993.

9.   Sanchez, R.; Barreto, J. Genomic abelian finite groups. Available online: https://doi.org/10.1101/2021.06.01.446543 (accessed on 1 March 2023).

10.   Frappat, L.; Sciarrino, A.; Sorba; P. Crystalizing the Genetic Code. *J. Biol. Phys.* **2001**, *27*, 1–34.

11. Sanchez, R.; Mackenzie, S. A. On the thermodynamics of DNA methylation process. *Scient. Rep.* **2023**, *13*, 8914.

12. Bessonov, N.; Butuzova, O.; Minarsky, A.; Penner R.; Soulé, C; Tosenberger, A.; Morozova, D. Morphogenesis software based on epigenetic code concept. *Comp. Struct. Biotech. J.* **2019**, *17*, 1203–1216.

13. Planat, M.; Amaral, M. M; Irwin, K. Algebraic morphology of DNA–RNA transcription and regulation. *Symmetry* **2023**, *15*, 770.

14. Schrödinger, E. *What Is Life? The Physical Aspect of the Living Cell*; Cambridge University Press: **1944**, Cambridge, UK.

15. 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**, *3*, 13290.

16. Grothendieck, A. *Esquisse d'un programme* **1984**, in Geometric Galois Actions I: Around Grothendieck's Esquisse D'un Programme, London Mathematical Society Lecture Note Series, vol. 242, Cambridge University Press Schneps and Lochak **1997**, pp. 243-283. Avalaible online: http://matematicas.unex.es/∼navarro/res/esquisseeng.pdf.

17. TATA Box: Available online: https://en.wikipedia.org/wiki/TATA_box (accessed on 1 September 2021).

18. Polyadenylation: Available online: https://en.wikipedia.org/wiki/Polyadenylation (accessed on 1 May 2023).

19. 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.

20. Goldman, W.M. Trace coordinates on Fricke spaces of some simple hyperbolic surfaces. *Eur. Math. Soc.* **2009**, *13*, 611–684.

21. Ashley, C.; Burelle J.P.; Lawton, S. Rank 1 character varieties of finitely presented groups. *Geom. Dedicata* **2018**, *192*, 1–19.

22. Jacob, W. F. Jacob, Santer, M., Dahlberg, A. E. A single-base change in the Shine-Dalgarno region of 16 rRNA of Escherichia colo affects translation of many proteins. *Proc. Natl. Acad. Sci. USA* **1987**, *84* 4257–4761.

23. Kozak, M. The scanning model for translation: an update. *J. Cell Biology* **1989**, *108* 229–241.

24. Kozak consensus sequence: Available online:
    https://en.wikipedia.org/wiki/Kozak_consensus_sequence (accessed on 1 January 2023).

25. Rittaud, B. On the average groth of random Fibonacci sequences. *J. Int. Seq.* **2007**, *10* Article 07.2.4.

26. microRNA. Available online: https://en.wikipedia.org/wiki/MicroRNA (accessed on 1 September 2022).

27. Fang, Y.; Pan, X.; Shen, H. B. Recent deep learning methodology development for RNA–RNA interaction prediction. *Symmetry* **2022**, *14*, 1302.

28. Medley, C. M.; Panzade G.; Zinovyeva, A. Y. MicroRNA strand selection,: unwinding the rules. *WIREs RNA* **2021**, *12*, e1627.

29. Dawson, O.; Piccinini, A. M. miR-155-3p: processing by-product or rising star in immunity and cancer? *Open Biol.* **2022**, *12*, 220070.

30. Kozomara, A.; Birgaonu, M.; Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucl. Acids Res.* **2019**, *47*, D155–D162.

31. miRBase: the microRNA database. Available online: https://www.mirbase.org/ (accessed on 1 November 2022).

32. Fromm, B.; Billipp, T.; Peck. L. E.; Johansen, M.; Tarver, J. E.; King, B. L.; Newcomb, J. M.; Sempere, L. F.; Flatmark, K.; Hovig, E.; Peterson, K. J. A uniform system for the annotation of human microRNA genes and the evolution of the human microRNAome. *Annu. Rev. Genet.* **2015**, *23*, 213–242.

33. He, Y.; Cai, Y.; Paii, P. M.; Ren, X.; Xia, Z. The Causes and Consequences of miR-503 Dysregulation and Its Impact on Cardiovascular Disease and Cancer. *Front. Pharmacol.* **2021**, *12*, 629611.

34. Planat, M.; Chester, D.; Amaral, M.; Irwin K. Fricke topological qubits. *Quant. Rep.* **2022**, *4*, 523-532.

35. Planat, M.; Amaral, M. M.; Chester, D.; Irwin, K. $SL(2,\mathbb{C})$ scheme processsing of singularities in quantum computing and genetics. *Axioms* **2023**, *12*, 233.

36. Mir-132. Available online: https://fr.wikipedia.org/wiki/Micro-ARN_7 (accessed on 1 June 2023).

37.   Sonkoly, E.; Stahle, M. Pivarsci, A. MicroRNAs and immunity: novel players in the regulation of normal
       immune function and inflammation. *Sem. in Cancer Biol.* **2008**, *18*, 131–140.
38.   Micro-ARN 7. Available online: https://en.wikipedia.org/wiki/MiR-132 (accessed on 1 June 2023).