Original Article

Group Theory of Messenger RNA Metabolism and Disease

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

[1]*Université de Franche-Comté, Institut FEMTO-ST CNRS UMR 6174, Besançon, France;* [2]*Quantum Gravity Research, Los Angeles, CA, USA*

Abstract

Background an objectives: 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 (mRNA) metabolism, showcasing the power of our approach.

Methods: To achieve this, we used: (a) infinite (finitely generated) groups , with generators representing the 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 microRNA seed in item (iii)]; (b) aperiodicity theory, which connects healthy groups to free groups of rank r and their profinite completion , and (c) the representation theory of groups over the space-time-spin group $SL_2(C)$, highlighting the role of surfaces with isolated singularities in the character variety.

Results: We investigate (1) mRNA translation in prokaryotes and eukaryotes, (2) polyadenylation in eukaryotes, which is crucial for nuclear export, translation, stability, and splicing of mRNA, (3) microRNAs involved in RNA silencing and post-transcriptional regulation of gene expression, and (4) identification of disrupted sequences that could lead to potential illnesses.

Conclusion: Our approach could potentially contribute to the understanding of the molecular mechanisms underlying various diseases and help develop new diagnostic or therapeutic strategies.

Introduction

Genome-scale metabolic pathways, genome-environment interactions, the immune response, post-transcriptional regulatory mechanisms, and oncohistones represent aspects of a research field connecting the heritable genetic code to other biological codes.[1–6] The aforementioned genetic code is defined precisely as a noninjective map from the 64 codons to the 20 amino acids. Both finite groups and quantum groups have leading roles in modeling this code.[7–10] More explicitly, according to Planat *et al.*,[8] complete quantum information is encoded in the 22 irreducible characters of the small group $(240,105) \cong Z_5 \rtimes 2O$, with $2O$ the binary octahedral group. The characters are put in correspondence with the DNA multiplets encoding the proteinogenic amino acids and the multiplicity is reflected in the dimension of the character representation. Further developments were explored in another study

by Planat *et al.*,[11] which showed that the small group $(336,118) \cong Z_7 \rtimes 2O$ is another model of the genetic code reflecting the symmetry of the Lsm–7 complexes in the spliceosome. The eight-fold symmetric histone complex was subsequently investigated by Planat *et al.*,[12] with the character table of the group $(384, 5,589) \cong Z_8 \rtimes 2O$.

The latest studies were the first to describe the role of a specific algebraic surface, called the Kummer surface, in the quantum modeling of the genetic code. From then on, we refer to the epigenetic code as all processes that reveal and execute gene expression. This includes DNA methylation processes,[13] messenger RNA (mRNA) translation preparation, the poly(A) tail, the RNA-induced silencing complex, a vital tool in gene regulation comprising single strands of RNA and double strands of small interfering RNA, and other regulatory nucleotide sequence fragments that are discarded after splicing. Ultimately, this involves a relation between the epigenetic code and morphogenesis.[14]

Chemical modifications of RNA also drive the metabolism of transcription of the genetic information. Post-transcriptional regulation of gene expression is a hot topic known as epitranscriptomics. There are more than 170 known types of RNA methylation processes but the most common in eukaryotes is the possible methylation of N^6-methyladenosine (m^6A) on sites with a specific short sequence $RR\underline{A}CH$ ($R = A$ or G, $H = A$, U, or C).[15–17]

To study the epigenetic code (hereinafter referred to as the e-code), we used infinite (finitely generated) groups denoted by f_p, and their representations over the (2 × 2) matrix group $SL_2(C)$, where the entries are complex numbers.[18,19] The significance of this group extends across all fields of physics, as it represents a space-time-spin group. In this study, we applied a mathematical field known as algebraic geometry to define the e-code, which has not been done before.

Our key 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 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 $SL_2(C)$ representations of f_p, termed the character variety, and specifically its basis called a Groebner basis G is necessary. The G comprises a set of surfaces. A surface within G containing isolated singularities indicates another potential disease that can be identified specifically, *e.g.*, relating to an oncogene or a neurological disorder.[19] 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 the study of Planat *et al.*,[19] is their connection to aperiodicity. Schrodinger proposed the periodicity of living crystals.[20] Planat *et al.*[19] characterized aperiodic DNA sequences.[21] We advanced 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 nucleotides) and $\hat{F}_2$ (corresponding to sequences containing three distinct nucleotides) have been examined in a prominent algebraic geometry treatise.[22] We present the details below in a manner that is accessible to a non-specialist reader. In the Methods section, we illustrate our mathematical concepts through a few simple pedagogical examples. In the Results section, we apply these concepts to cases of mRNA translation, microRNAs (miRNAs), and m^6A methylation. In the Discussion, we provide additional comments, a summary diagram, and perspectives.

Methods and preliminary results

Infinite finitely generated groups f_p and free groups F_r

TATA box

We 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 noncoding 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 (https://en.wikipedia.org/wiki/TATA_box).

In our methodology, we defined the group $f_p = \langle A,T|rel\rangle$, which contains an infinite number of 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 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···]. Interestingly, the group $H_3 = \langle A, T|A^2 = T^3\rangle$) has 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 $PSL(2,Z)$ – the projective special linear 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 as the card seq of both groups is the same, but we can easily verify that f_p and H_3 are not isomorphic. According to Planat *et al.*,[23] the 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 ≠ 3 or q ≠ 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 (https://en.wikipedia.org/wiki/Poly-adenylation). Polyadenylation involves the addition of a poly(A) tail to an RNA transcript, usually a mRNA. A consensus poly(A) sequence takes the form rel1 = AAUAAA, which corresponds to a two-generator group of the form $f_p = \langle AU|rel1\rangle$. The card seq of such a group is found to be [1,1,1,1,1,1,1,1,1,1,···], 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 neither group is isomorphic. Another consensus poly(A) sequence takes the form rel2 = UGUAA, which corresponds to a three-generator group of the form $f_p \langle A, U, G|rel2\rangle$. The card seq of such a group is found to be [1,3,7,26,97,624,4,163,···]. 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.[19,21,23] 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.

Aperiodic sequences, their attached groups f_p and free groups

In this subsection, we 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.[21,23] Consider the motif rel = *TTTATTA*, which serves as a consensus sequence for the transcription factor of the DBX gene in *Drosophila melanogaster* (fruit fly). This gene is involved in neuronal specification and differentiation. The group $f_p = \langle A, T|rel\rangle$ has the same card seq as the free group F_1 of rank 1. Furthermore, by splitting rel into two segments rel = $rel_A rel_T$ and applying the substitution maps $A \rightarrow rel_A = TTTA$, $T \rightarrow rel_T = TTA$, we generate the substitution sequence $S_{DBX} = A,T,AT,TTTATTA,TTATTATTATT$ *TATTATTATTTA*,···. On inspection, it is straightforward to observe that all finitely generated groups $f_p^{(l)}$, with their generators being *AT,TTTATTA,TTATTATTATTTATTATTATTTA*,···, respectively, have the card seq of F_1.

As per the findings of Planat *et al.*,[23] for a substitution rule to be considered aperiodic it 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 $\gg$ 0. (2) The Perron–Frobenius λ_{PF} eigenvalue must be irrational. It is 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 as $M^2 \gg 0$ and $\lambda_{PF} = 3 + \sqrt{13}/2$. Conditions (1) and (2) are satisfied, implying that the substitution S_{DBX} is aperiodic. Of note, numerous other genes have transcription factors with a motif rel generating an aperiodic sequence.[21]

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 question of why the aforementioned groups $f_p(l)$ produce the same card seq as that of the free group F_r. The tentative answer is that the profinite completion of all groups $f_p(l)$ is the profinite group $\hat{F}_1$. By making this observation, we aligned the aperiodicity of sequences with profinite groups. Profinite groups were introduced by Grothendieck in the context of algebraic geometry.[22] 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 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 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 analog of each of these normal subgroups. The profinite group $\hat{G}$ exhibits several properties: it is nonabelian, residually finite, (meaning that for any nonidentity 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 too complex to handle.

Considering the profinite group $\hat{F}_1$, we 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, we 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^*, as it corresponds to the elements of Z_p with a valuation of zero. The p-adic integers are a special class of numbers used in number theory and algebraic geometry.

Considering the profinite group $\hat{F}_2$, we briefly discuss $\hat{F}_2$. This topic was first described by Grothendieck.[22] The subject is complex and connected to the so-called Belyi theorem, a fundamental result that establishes a connection between algebraic curves de-

fined over the algebraic closure of the rationals, Q, and certain rational functions called Belyi functions. An algebraic curve defined over Q can be represented as a branched covering of the Riemann sphere (the complex projective line $P^1(C)$) 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 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:

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

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

$SL_2(C)$ representations of groups f_p and a Groebner basis G

While the previous section describing 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 $SL_2(C)$, as we did in previous studies.[18,19,21] Representations of f_p in $SL_2(C)$ are homomorphisms $\rho: f_p \rightarrow SL_2(C)$ 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 used to determine an algebraic set by taking the quotient of the set of representations ρ by the group $SL_2(C)$, which acts by conjugation on representations.[24,25] In such papers, we showed 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 $G(X)$, whose factors define hypersurfaces.

Our previous paper focused on a possible algebraic approach of topological quantum computing.[18] In two subsequent papers,[19,21] we investigated $SL_2(C)$ representations of short DNA/RNA sequences (*e.g.*, the consensus sequence of a transcription factor or the seed of a miRNA) and related them to a potential disease. 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.[19] Often encountered surfaces are degree p Del Pezzo surfaces where $1 \le p \le 9$. A rational surface may either be nonsingular, almost nonsingular, having only isolated singularities, or singular. Almost nonsingular surfaces are key 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, $S^{(lA1, mA2, nA3, \cdots)}$ 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 $S^{(4A1)} = xyz + x^2 + y^2 + z^2 - 4$.[19]

For a three-generator group f_p, the factors of $G(X)$ are seven-dimensional surfaces of the form $S_{a,b,c,d}(x,y,z)$. Some of them belong to the Fricke family,[19] 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 $G(X)$ are 14-dimensional surfaces containing four copies of the form $S(x,y,z)$, $S(x,u,v)$, $S(y,u,v)$, and
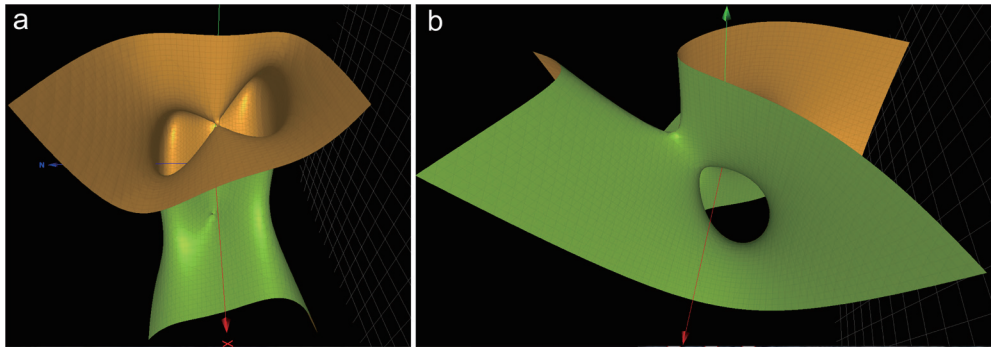
Fig. 1. Two types of Del Pezzo surfaces. (a) Degree 2 Del Pezzo surface within G_{TATA}. (b) Degree 3 Del Pezzo surface $S^{(A1)}$ within G_{rel1}.

$S(z,v,w)$ for selected choices of eight parameters.

Groebner basis of the TATA box

The Groebner basis for the character variety associated with the f_p group of generators rel = TATAAAA of the TATA box as discussed above, is found to be:

$$G_{TATA} = (z^4 - xy^2 - xyz + x^2 + y^2 + yz - 3z^2 + x - 2) \, (x^2z - xy - xz + y - z) \, S^{(A2)} S^{(A4)} \, (x^3 - z^2 - 3x + 2),$$

where $S^{(A2)} = x^2y - z^3 - xz - y + 3z$ and $S^{(A4)} = xz^2 - x^2 - yz - x + 2$ are degree 3 Del Pezzo surfaces. The Groebner basis G_{TATA} comprises a degree 2 Del Pezzo surface (Fig. 1a, 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).

Groebner basis for polyadenylation signals

For the first polyadenylation signal considered in the paragraph describing infinite finitely generated groups. The relation of the f_p group is rel1 = AAUAAA. The corresponding Groebner basis is:

$$G_{rel1} = 3 \text{ rational scrolls} \times P^2 \times S^{(4A1)} S^{(A1)} \times \text{curve}.$$

The Groebner basis G_{rel1} contains three rational scrolls, a surface birationally equivalent to the projective plane P^2, the Cayley cubic $S^{(4A1)}$, the degree 3 Del Pezzo surface $S^{(A1)} = x^2y - xz^2 - xz + yz + x - y$ (Fig. 1b) and a curve.

For the second polyadenylation signal considered above in the paragraph describing groups f_p and F_r, the relation of the f_p group is rel2 = UGUAA. The factors of $G(X)$ are seven-dimensional hypersurfaces $S_{a,b,c,d}(x,y,z)$. However, by choosing specific parameters, such as $S_{0,0,0,0}(x,y,z)$ or $S_{1,1,1,1}(x,y,z)$, we obtained three-dimensional surfaces. These were found to be degree 3 Del Pezzo surfaces with simple singularities of the form $S^{(lA2)}$, with l = 1, 2, or 3, quadrics, or curves.

Groebner basis of the transcription factor of DBX gene

For the *DBX* gene studied in the paragraph on aperiodic sequences, the Groebner basis takes the form of G_{DBX} = scroll $\times P^2 \times S^{(A4)} \times S^{(A2)} \times S^{(4A1)} \times$ 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 $S^{(A4)} = yz^2 - y^2 - xz - y^2$, $S^{(A2)} = z^3 - xy^2 + yz + - 3z$, the Cayley cubic $S^{(4A1)}$ and curve = $y^3 - z^2 - 3y + 2$.

Further results

In this section, we describe additional results related to mRNA me-

tabolism and miRNA.

Algebraic geometry of mRNA translation

Shine-Dalgarno box

Ribosomal RNA is a type of noncoding RNA and 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 mRNA by the ribosome. A major factor in this recognition is an mRNA–ribosomal RNA interaction first proposed by Shine and Dalgarno.[26] They proposed that the ribosomal nucleotides recognize the complementary purine-rich sequence rel3 = AGGAGGU, which is found approximately eight 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 as that of the free group F_2. The $SL_2(C)$ character variety is a scheme X whose Groebner basis $G(X)$ comprises 7-dimensional surfaces $S_{a,b,c,d}(x,y,z)$. By projecting to three dimensions, one gets surfaces like $S_{0,0,0,0}(x,y,z)$ and $S_{1,1,1,1}(x,y,z)$ as in the paragraph describing $SL_2(C)$ representations of groups f_p. We find degree 3 Del Pezzo surfaces with isolated singularities $S^{(A1)} = x^2y + yz^2 + 4xz + 4y$ and $x^2y + yz^2 + x + z^2 + 6xz + 5y - 6z - 7$, $S^{(A2)} = xyz + 2x^2 + z^2 + 4$ and $S^{(A4)} = xyz + 3x^2 + z^2 - 5z$, 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.[27] The small (40S) subunit of eukaryotic ribosomes bind, initially at the capped 5′-end of the mRNA 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 Kozak (https://en.wikipedia.org/wiki/Kozak) consensus sequence 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 as 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 the paragraph describing aperiodic sequences. By splitting rel4 into four segments rel4 = $rel_A rel_C rel_G rel_U$ and applying the substitution maps $C \rightarrow rel_C = A$, $A \rightarrow rel_A = CCAUG$, $U \rightarrow rel_U = G$, $G \rightarrow rel_G = C$, we generated the substitution se-

quence: $S_{\text{Kozak}} = C,A,U,G,CAUG,ACCAUGGC,CCAUGA^2CCAU$ $GGC^2A,\cdots$.

On inspection, it is straightforward to observe that all finitely generated groups $f_p^{(l)}$ with their generations being $CAUG$, AC-$CAUGGC$, $CCAUGA^2CCAUGGC^2A,\cdots$, respectively, have a 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 as $M^4 \gg 0$ and $\lambda_{PF} \approx 2.2055694$ is the only real (and irrational) solution of the equation $x^3 - 2x^2 - 1 = 0$. Conditions (1) and (2) for aperiodic sequences are satisfied, implying that the substitution S_{Kozak} is aperiodic. Rittaud discussed the connection of the later Perron–Frobenius eigenvalue to random Fibonacci sequences.[28]

Mutation of a purine at position −3 with respect to the AUG codon is known to be associated with diseases including a type of thalassemia owing to a bad initiation of alpha-globin.[27] 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 card seq 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.

Algebraic geometry of miRNAs

miRNAs are small, single-stranded, noncoding RNA molecules 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(https://en.wikipedia.org/wiki/MicroRNA).[29] 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 precursor (pre)-miRNA. Pre-miRNAs are approximately 70 nucleotides long and fold into imperfect stem-loop structures. A miRNA consists of a duplex comprising two strands (−5p and −3p). However, a single strand is selected into the RNA-induced silencing complex to serve as a template during the transcription of a complementary mRNA.[30,31] For details of the miRNA sequences, we use the Mir database (https://www.mirbase.org/).[32,33] It should be emphasized that a given miRNA may have hundreds of different mRNA targets and a single target may be regulated by multiple miRNAs. For previous discussions of how to define an f_p group from the seed of a miRNA, the reader may consult a recent review.[19] Below, we focus on other examples.

miRNA hsa-mir-122

mir-122 is a tissue-specific miRNA that is highly expressed in the liver.[34] It is involved in cholesterol accumulation and fatty acid metabolism. It has a leading role in controlling the hepatitis C virus.[35,36] The seed region for mir-122-5p is seed0 = GGAGUGU. The corresponding group $f_p = \langle C, G, U|\text{seed0}\rangle$ has the card seq of the free group F_2. Let us first check if the seed sequence is aperiodic. By splitting seed0 into three segments seed0 = seed$_A$ seed$_G$ seed$_U$ and applying the substitution maps $A \to$ seed$_A =$ GG, $G \to$ seed$_G =$ AGU, $U \to$ seed$_U =$ GU, one can check that the finitely generated groups $f_p^{(l)}$ with generators $GGAGUGU$, $AGUAGUG$-$GAGUGUAGUGU$, possess the card seq of the free group F_2.

Following the method described in the section on aperiodic sequences, their attached and free groups, one gets the (primitive) substitution matrix:

$$M = \begin{pmatrix} 0 & 1 & 0 \\ 2 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

whose characteristic polynomial $\lambda^3 - 2\lambda^2 - 2\lambda + 2$ has three real roots. The largest one is the (irrational) Perron–Frobenius eigenvalue $\lambda_{PF} \approx 2.481194$. One concludes that the sequence seed0 is aperiodic.

Let us now look at the Groebner basis for the $SL_2(\mathbb{C})$ representation of f_p with the method described above. One obtains:

$G_{\text{mir-122−5p}}(0,0,0,0) = 8yz(2 - z^2)$ and $G_{\text{mir-122−5p}}(1,1,1,1)$
$= -4\, z^2(\text{x} - z^2 + z + 1)\,(y + z^3 - z^2 - 2z)$

One can check that all values of the parameters $Ga,b,c,d\,(x, y, z)$ only contain factors that are curves and not surfaces.

miRNA hsa-mir-503

The slowest evolving miRNA gene in the human species (hsa) is hsa-mir-503 (https://www.mirbase.org/). It regulates gene expression in various pathological processes of diseases, including carcinogenesis, angiogenesis, tissue fibrosis, and oxidative stress.[37] The seed region of mir-503-5p is seed1 = AGCAGCGG. The corresponding group $f_p = \langle A, C, U|\text{seed1}\rangle$ 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: $G_{mir-503-5p}(0,0,0,0)$ $= S^{(4A1)}(x,y,z)$, which is the already mentioned Cayley cubic. For $(a,b,c,d) = (1,1,0,0)$, $G_{mir-503-5p}(1,1,0,0) = -3xyz\kappa_3(x,y,z)$, where $\kappa_3(x,y,z)$ is the Fricke surface described by Planat *et al*.[38] For $(a,b,c,d) = (1,1,1,1)$, there are several more polynomials. One of which defines the Fricke surface $xyz + x^2 + y^2 + z^2 - 2x - y - 2 = 0$. 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 has a role in neuropathogenesis. The considered seed region for hsa-mir-146a-5p is seed2 = GAGAAC (https://www.mirbase.org/). 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 $G_{\text{hsa-146a−5p}}(0,0,0,0) = (xz + y + 2)\,(y - z^2 + 2)^2\,(x^2 + z^2 - 2y - 4)\,S^{(3A2)}$, where $S^{(3A2)} = z^3 - xy - 2yz - 2x - 4z$. The Groebner basis with parameters $(a,b,c,d) = (1,1,1,1)$ is of the form $G_{\text{hsa-146a−5p}}(1,1,1,1) = DP^4 \times f^{(2A2)}\times$ quadric $\times$ curves, where DP^4 is a degree 4 del Pezzo surface.

miRNAs and disease

As described previously,[19] a potential disease is associated with f_p groups that fail to satisfy at least one of three requirements: (1) the card seq of f_p should be that of a free group F_r; (2) the generating sequence should be aperiodic; or (3) the $SL_2(\mathbb{C})$ character variety of f_p should have a Groebner basis devoid of isolated singularities even though the f_p group may have the card seq of a free group.[19] Following these criteria, the sequence hsa-mir-122-5p is healthy but the sequences hsa-mir-503-5p and hsa-mir-146a-5p are not because criterion three is not satisfied. Additional examples can be found in our previous study.[19]

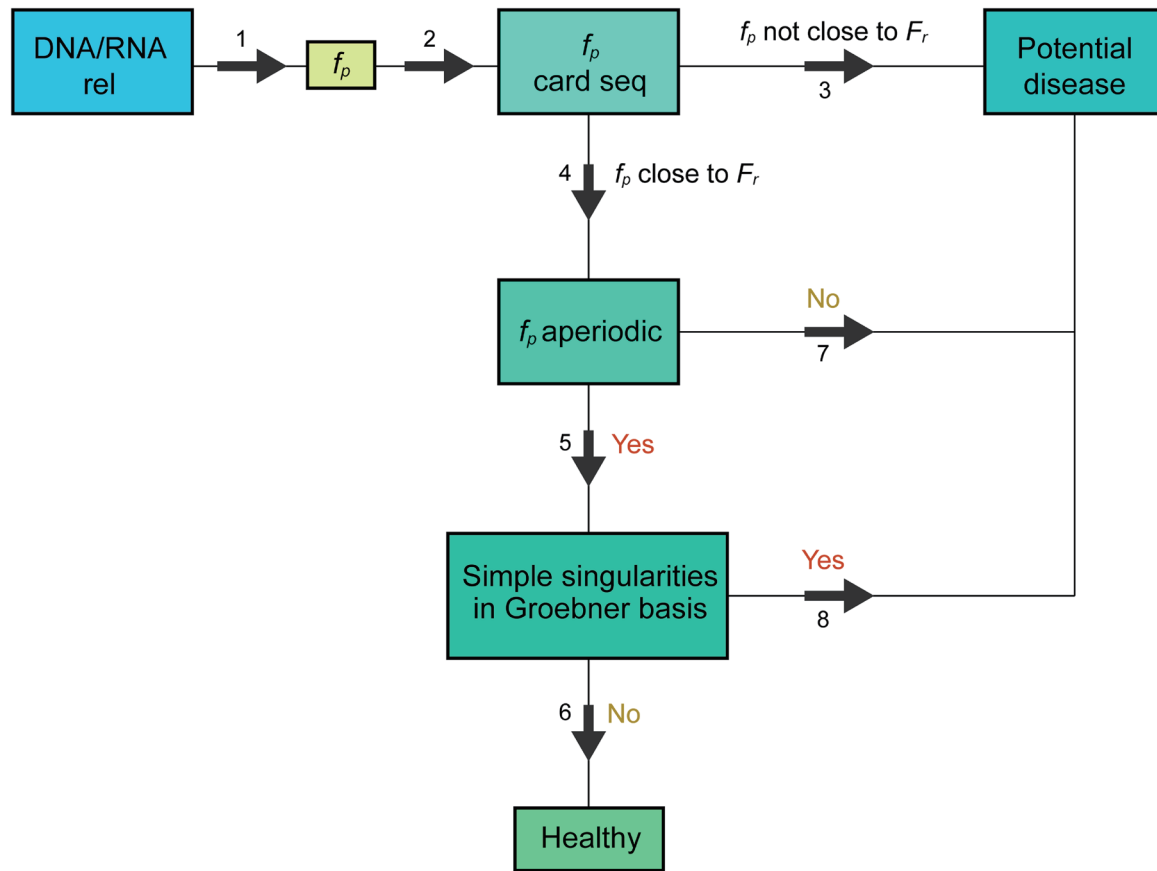In addition to isolated singularities, the Groebner basis may

Fig. 2. Diagram of the main results discussed in the text. For example, for the transcription factor of the gene *EGR1*, rel = GCGTGGGCG [25, Section 4.1.2], the path is 1 → 2 → 4 → 5 → 6 showing no risk of disease. But for the transcription factor of gene *DBX* (see the subsections about aperiodic sequences and the SL_2(C) representations of groups), rel= TTTATTA, the path is 1 → 2 → 4 → 5 → 8 meaning a potential disease (see Table 1).

contain unique surfaces that are not simply singular. The DP^4 surface in $G_{\text{hsa-146a-5p}}(1,1,1,1)$ is an example of a singular surface. Further mathematical evaluation is required to investigate these surfaces.[39] However, we will not include them in this review.

Discussion

Figure 2 summarizes our key results. Given a short DNA/RNA sequence, rel that represents a consensus sequence in a transcription factor, the seed of a miRNA, or a relevant sequence in mRNA recognition and processing, we constructed a finitely generated group, f_p. The architecture of subgroups, card seq, within this group was computed, as described in the subsection about the infinite finitely generated groups f_p. If the f_p card seq matches that of the free group F_r (of rank r = nt − 1), we proceed to path four; otherwise, a potential disease could be in sight (path three). After reaching path four, the next step involves checking the aperiodicity of rel and the corresponding f_p group, as described in the subsection about aperiodic sequences and their attached groups f_p. The final step is to examine the presence (or absence) of isolated singularities in the Groebner basis G for the SL_2(C) character variety associated with f_p, as outlined in the subsection about SL_2(C) representations of groups f_p. For a healthy sequence, the path concludes at six, while a potential disease may be indicated if the path ends at three, seven, or eight.

In Table 1, we provide several examples of paths.[23,31,36,37,40] 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 four distinct nucleotides, like the sequence of transcription factor FOX or the Kozak sequence rel4, it is difficult to make a conclusion about the risk of a disease. The generic Groebner basis[1] $G(x,y,z)$ always contains a surface with isolated singularities such as $S^{(4A1)}$ and $S^{(3A1)}$ and there are four copies of them. The termination {6,8} applies for this case.

Algebraic geometry of m⁶A modifications

As mentioned in the Introduction, a subfield of epigenetics deals with post-transcriptional mRNA modifications. m^6A is the most frequent modification in most eukaryotes. But m^6A is also present in bacteria, with the consensus motif $GCCAG$.[41,42] An interesting aspect is that the mRNA m^6A motif in bacteria is distinct from the consensus motif in eukaryotes (RRACH). This features the evolutionary machinery present in the last eukaryotic common ancestor compared to the last universal common ancestor.[43] In Table 2, we provide details of the group generated by these sequences, when the sequence is aperiodic and/or has a Groebner basis of its character variety containing an isolated singularity. The path in the diagram of Figure 2 is shown in Table 1.

Table 1. A few possible paths in the Figure 2 diagram that terminates at path six (healthy) or three, seven, or eight (potential disease)

Sequence	rel	Path
EGR1[23]	GCGTGGGCG	1→2→4→5→6
FOS[23]	TGAGTCA	1→2→4→5→{6,8}
Nanog[23]	TAATGG	1→2→4→{7,8}
DBX	TTTATTA	1→2→4→5→8
TATA	TATAAAA	1→2→3→(7,8)
Poly(A) (rel1)	AAUAAA	1→2→3→{7,8}
Poly(A) (rel2)	UGUAA	1→2→4→{7,8}
Shine-Dalgarno (rel3)	AGGAGGU	1→2→4→5→8
Kozak (rel4)	ACCAUGGC	1→2→4→5→{6,8}
Kozak (rel4')	CCCAUGGC	1→2→4→7
hsa-mir-122-5p[36](seed0)	GGAGUGU	1→2→4→5→6
hsa-mir-132-5p (https://fr.wikipedia.org/wiki/Micro-ARN_7)	CCGUGGC	1→2→4→5→6
mir-503-5p (seed1)[37]	AGCAGCGG	1→2→5→8
mir-146a-5p (seed2)[40]	GAGAAC	1→2→{7,8}
hsa-mir-7-5p (https://en.wikipedia.org/wiki/MiR-132)	GGAAGA	1→2→{3,7,8}
hsa-mir-7-5p	GGAAGAC	1→2→4→5→6
hsa-mir-7-3p	AACAAAU	1→2→4→7
hsa-mir-155-3p[31,40]	UCCUAC	1→2→4→{7,8}
hsa-mir-155-3p	UCCUACA	1→2→3

The set {6,8} denotes a lack of a clear conclusion of the existence of an isolated singularity. The selected examples are displayed in three parts that are transcription factors (first group), regulating elements in introns (second group) and miRNAs (third group). Details are given in the text. Otherwise a reference is provided.

Only the bacterial sequence leads to a path terminating at edge 6 of the diagram of Figure 2. In the closest eukaryotic sequence GGACA (from the viewpoint of group analysis), isolated singularities are found, such as the degree 3 Del Pezzo surface $S^{(A2A2)}$ = $y^3 - 2xz - 4y$. The other sequences are not aperiodic. From the biological point of view, it is known that an appropriate level of m^6A methylation is beneficial, but it may be a risk to drive it in an artificial way because it may destroy the delicate balance of regu-

lations performed within the mRNA.

Conclusions

Our approach was 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

Table 2. Detailed group theoretical analysis of m^6A modifications for bacteria (the sequence GCCAG) and eukaryotes (sequence $RR\underline{A}CH$ (R = A or G, H = A, U, or C))

Sequence	Group	Aperiodic	Groebner basis	Path
Bacterial				
GCCAG	F_2	1.83928	No	1→2→4→5→6
Eukaryote				
AAACA	F_1	No		1→2→4→{7,8}
AAACC	H_3	No		1→2→{3,7}
AAACU	F_2	No	$S^{(A2)}$, $S^{(A1A2)}$ No	1→2→4→7
GGACA	F_2	1.83928	No	1→2→4→5→8
GGACC	F_2	No	$S^{(A2)}$, $S^{(A2A2)}$ No	1→2→4→7
GGACU	F_3	No	Unknown	1→2→4→7

Column 2 is the group closer to the f_p group generated by the sequence in column 1 (F_r is for the free group of rank r, H_3 is for the modular group $PSL(2,Z)$. If the sequence is aperiodic, the Perron–Frobenius eigenvalue λ_{PF} is given in column 3. The type of isolated singularity, if any, is in column 4. The path in the diagram of Figure 2 is shown in column 5.

that reveals and executes gene expression, has a sophisticated structure that our mathematical approach aimed to elucidate.

Conflict of interest

The authors declare that they have no conflicts of interest.

Author contributions

Conceptualization (MP, FF, KI), methodology (MP, DC, RA), software (MP), validation (RA, FF, DC, MMA), formal analysis (MP, MMA), investigation (MP, DC, FF, MMA), writing and original draft preparation (MP), writing, review and editing (MP) visualization (FF, RA), supervision (MP, KI), project administration (KI), and funding acquisition (KI). All authors have read and approved the final version of the manuscript.

Data share statement

Computational data are available from the authors upon reasonable request.

References

[1] Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. Current status and applications of genome-scale metabolic models. Genome Biol 2019;20(1):121. doi:10.1186/s13059-019-1730-3, PMID:31196170.

[2] Romão L. mRNA metabolism in health and disease. Biomedicines 2022;10(9):2262. doi:10.3390/biomedicines10092262, PMID:36140363.

[3] Peedicayil J. Genome-environment interactions and psychiatric disorders. Biomedicines 2023;11(4):1209. doi:10.3390/biomedicines11041209, PMID:37189827.

[4] Scharf S, Ackermann J, Bender L, Wurzel P, Schäfer H, Hansmann ML, *et al*. Holistic view on the structure of immune response: petri net model. Biomedicines 2023;11(2):452. doi:10.3390/biomedicines11020452, PMID:36830988.

[5] Marques AR, Santos JX, Martiniano H, Vilela J, Rasga C, Romão L, *et al*. Gene variants involved in nonsense-mediated mrna decay suggest a role in autism spectrum disorder. Biomedicines 2022;10(3):665. doi:10.3390/biomedicines10030665, PMID:35327467.

[6] Wan YCE, Chan KM. Histone H2B mutations in cancer. Biomedicines 2021;9(6):694. doi:10.3390/biomedicines9060694, PMID:34205231.

[7] Fimmel E, Giannerini S, Gonzalez DL, Strüngmann L. Circular codes, symmetries and transformations. J Math Biol 2015;70(7):1623–1644. doi:10.1007/s00285-014-0806-7, PMID:25008961.

[8] Planat M, Aschheim R, Amaral MM, Fang F, Irwin K. Complete quantum information in the DNA genetic code. Symmetry 2020;12:1993. doi:10.3390/sym12121993.

[9] Sanchez R, Barreto J. Genomic abelian finite groups. bioRxiv [Preprint] 2023. doi:10.1101/2021.06.01.446543.

[10] Frappat L, Sciarrino A, Sorba P. Crystalizing the genetic code. J Biol Phys 2001;27(1):1–34. doi:10.1023/A:1011874407742, PMID:23345731.

[11] Planat M, Chester D, Aschheim R, Amaral MM, Fang F, Irwin K. Finite groups for the Kummer surface: the genetic code and quantum gravity. Quantum Rep 2021;3:68–79. doi:10.3390/quantum3010005.

[12] Planat M, Aschheim R, Amaral MM, Fang F, Irwin K. Quantum information in the protein codes, 3-manifolds and the Kummer surface. Symmetry 2021;13:1146. doi:10.3390/sym13071146.

[13] Sanchez R, Mackenzie SA. On the thermodynamics of DNA methylation process. Sci Rep 2023;13(1):8914. doi:10.1038/s41598-023-35166-9, PMID:37264042.

[14] Bessonov N, Butuzova O, Minarsky A, Penner R, Soulé C, Tosenberger A, *et al*. Morphogenesis software based on epigenetic code concept. Comput Struct Biotechnol J 2019;17:1203–1216. doi:10.1016/j.csbj.2019.08.007, PMID:31666938.

[15] Vissers C, Sinha A, Ming GL, Song H. The epitranscriptome in stem cell biology and neural development. Neurobiol Dis 2020;146:105139. doi:10.1016/j.nbd.2020.105139, PMID:33065280.

[16] Wang S, Lv W, Li T, Zhang S, Wang H, Li X, *et al*. Dynamic regulation and functions of mRNA m6A modification. Cancer Cell Int 2022;22(1):48. doi:10.1186/s12935-022-02452-x, PMID:35093087.

[17] Widagdo J, Wong JJ, Anggono V. The m(6)A-epitranscriptome in brain plasticity, learning and memory. Semin Cell Dev Biol 2022;125:110–121. doi:10.1016/j.semcdb.2021.05.023, PMID:34053866.

[18] Planat M, Amaral MM, Fang F, Chester D, Aschheim R, Irwin K. Character varieties and algebraic surfaces for the topology of quantum computing. Symmetry 2022;14:915. doi:10.3390/sym14050915.

[19] Planat M, Amaral MM, Irwin K. Algebraic morphology of DNA-RNA transcription and regulation. Symmetry 2023;15:770. doi:10.3390/sym15030770.

[20] Schrödinger E, editor. What Is Life? The Physical Aspect of the Living Cell. Cambridge: Cambridge University Press; 1944.

[21] Planat M, Amaral MM, 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(21):13290. doi:10.3390/ijms232113290, PMID:36362076.

[22] Grothendieck A, editor. Lecture Series of the London Mathematical Society. Cambridge: Cambridge University Press; 1997:243–283. doi:10.1017/CBO9780511758874.003.

[23] Planat M, Amaral MM, 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(4):1417–1433. doi:10.3390/cimb44040095, PMID:35723353.

[24] Goldman WM. Trace coordinates on Fricke spaces of some simple hyperbolic surfaces. Eur Math Soc 2009;13:611–684. doi:10.4171/055-1/16.

[25] Ashley C, Burelle JP, Lawton S. Rank 1 character varieties of finitely presented groups. Geom Dedicata 2018;192:1–19. doi:10.1007/s10711-017-0281-6.

[26] Jacob WF, Santer M, Dahlberg AE. A single base change in the Shine-Dalgarno region of 16S rRNA of Escherichia coli affects translation of many proteins. Proc Natl Acad Sci U S A 1987;84(14):4757–4761. doi:10.1073/pnas.84.14.4757, PMID:2440027.

[27] Kozak M. The scanning model for translation: an update. J Cell Biol 1989;108(2):229–241. doi:10.1083/jcb.108.2.229, PMID:2645293.

[28] Rittaud B. On the average growth of random Fibonacci sequences. J Int Seq 2007;10:07.2.4.

[29] Fang Y, Pan X, Shen HB. Recent deep learning methodology development for RNA-RNA interaction prediction. Symmetry 2022;14:1302. doi:10.3390/sym14071302.

[30] Medley JC, Panzade G, Zinovyeva AY. microRNA strand selection: Unwinding the rules. Wiley Interdiscip Rev RNA 2021;12(3):e1627. doi:10.1002/wrna.1627, PMID:32954644.

[31] Dawson O, Piccinini AM. miR-155-3p: processing by-product or rising star in immunity and cancer? Open Biol 2022;12(5):220070. doi:10.1098/rsob.220070, PMID:35611569.

[32] Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res 2019;47(D1):D155–D162. doi:10.1093/nar/gky1141, PMID:30423142.

[33] Fromm B, Billipp T, Peck LE, Johansen M, Tarver JE, King BL, *et al*. A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. Annu Rev Genet 2015;49:213–242. doi:10.1146/annurev-genet-120213-092023, PMID:26473382.

[34] Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T, Pallasch C, *et al*. Distribution of miRNA expression across human tissues. Nucleic Acids Res 2016;44(8):3865–3877. doi:10.1093/nar/gkw116, PMID:26921406.

[35] Girard M, Jacquemin E, Munnich A, Lyonnet S, Henrion-Caude A. miR-122, a paradigm for the role of microRNAs in the liver. J Hepatol 2008;48(4):648–656. doi:10.1016/j.jhep.2008.01.019, PMID:18291553.

[36] Hu J, Xu Y, Hao J, Wang S, Li C, Meng S. MiR-122 in hepatic function and liver diseases. Protein Cell 2012;3(5):364–371. doi:10.1007/s13238-012-2036-3, PMID:22610888.

[37] He Y, Cai Y, Pai PM, 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. doi:10.3389/fphar.2021.629611, PMID:33762949.

[38] Planat M, Chester D, Amaral M, Irwin K. Fricke topological qubits. Quant Rep 2022;4:523–532. doi:10.3390/quantum4040037.

[39] Planat M, Amaral MM, Chester D, Irwin K. SL(2,C) scheme processs-ing of singularities in quantum computing and genetics. Axioms 2023;12:233. doi:10.3390/axioms12030233.

[40] Sonkoly E, Ståhle M, Pivarcsi A. MicroRNAs and immunity: novel players in the regulation of normal immune function and inflammation. Semin Cancer Biol 2008;18(2):131–140. doi:10.1016/j.semcancer.2008.01.005, PMID:18291670.

[41] Deng X, Chen K, Luo GZ, Weng X, Ji Q, Zhou T, *et al*. Widespread occurrence of N6-methyladenosine in bacterial mRNA. Nucleic Acids Res 2015;43(13):6557–6567. doi:10.1093/nar/gkv596, PMID:26068471.

[42] Gao R, Tsui PH, Wu S, Tai DI, Bin G, Zhou Z. Ultrasound entropy imaging based on the kernel density estimation: a new approach to hepatic steatosis characterization. Diagnostics (Basel) 2023;13(24):3646. doi:10.3390/diagnostics13243646, PMID:38132230.

[43] Liu C, Cao J, Zhang H, Yin J. Evolutionary history of RNA modifications at N6-adenosine originating from the R-M system in eukaryotes and prokaryotes. Biology (Basel) 2022;11(2):214. doi:10.3390/biology11020214, PMID:35205080.