Quantum Information in the Protein Codes, 3-manifolds and the Kummer Surface

Fang Fang, Michel Planat, Raymond Aschheim, Marcelo M Amaral, Klee Irwin

QUANTUM INFORMATION IN THE PROTEIN CODES, 3-MANIFOLDS AND THE KUMMER SURFACE

MICHEL PLANAT†, RAYMOND ASCHHEIM‡,
MARCELO M. AMARAL‡, FANG FANG‡ AND KLEE IRWIN‡

Abstract. Every protein consists of a linear sequence over an alphabet of 20 letters/amino acids. The sequence unfolds in the 3-dimensional space through secondary (local foldings), tertiary (bonds) and quaternary (disjoint multiple) structures. The mere existence of the genetic code for the 20 letters of the linear chain could be predicted with the (informationally complete) irreducible characters of the finite group $G_n := \mathbb{Z}_n \rtimes 2O$ (with $n = 5$ or 7 and $2O$ the binary octahedral group) in our previous two papers. It turns out that some quaternary structures of protein complexes display n-fold symmetries. We propose an approach of secondary structures based on free group theory. Our results are compared to other approaches of predicting secondary structures of proteins in terms of α helices, β sheets and coils, or more refined techniques. It is shown that the secondary structure of proteins shows similarities to the structure of some hyperbolic 3-manifolds. The hyperbolic 3-manifold of smallest volume –Gieseking manifold–, some other 3 manifolds and Grothendieck's cartographic group are singled out as tentative models of such secondary structures. For the quaternary structure, there are links to the Kummer surface.

Arxiv: quant-ph, math.GR, math.AG, q-bio.OT
PACS: 02.20.-a, 03.65.Fd, 82.39.Rt, 87.10.-e, 87.14.-g, 87.14.Ee
MSC codes: 20C15, 92D20, 20E45, 14H52, 14J28, 58D19, 57M27
Keywords: protein structure, DNA genetic code, informationally complete characters, finite groups, 3-manifolds, Kummer surface, cartographic group

1. Introduction

Proteins are long polymeric linear chains encoded with 20 amino acid residues arranged in a biologically functional way. Today the protein data base (or PDB) contain about 1.8×10^5 entries [1]. Proteins may perform a large variety of fonctions in living cells and organisms including molecular recognition, catalysing metabolic reactions, DNA replication and structural support for molecules. The sequence of aminoacids results in many different three-dimensional foldings that happen to be more conserved during evolution than the sequences themselves. The structure of proteins determines their biological function [2].

A coarse-grained representation of the backbone structure of the linear chain in a protein results into three main elements that are α helices and β pleated sheets, due to the interactions between atoms and backbones, and random coils that indicate an absence of a regular structure. The ordered structures are held in shape by hydrogen bonds, which form between the carbonyl of one amino acid and the amino of another. In an α helix, there is a pattern of bonds that puts the polypeptide chain into a helical structure

with each turn of the helix containing 3.6 amino acids [3]. In a β pleated sheet, two or more segments of a polypeptide chain line up next to each other, forming a sheet-like structure held together by hydrogen bonds [4]. The three main elements of a protein linear chain are usually denoted H (if the segments form an α helix), E (if the segments form a β pleated sheet) and C (if the segments form a coil) and constitute what is called the secondary structure of the protein.

In this paper, we are interested in the universality of the two- or three-letter (secondary) code found in proteins. The letters are segments of the protein that correspond to an α helix H, a β pleated sheet E or a random coil C. Our view of the connection of proteins as words with two letters (or three letters) and free group theory is as follows. One defines the two-letter group $G = \langle H, C|\text{rel}(H, C)\rangle$ or the three-letter group $G := \langle H, E, C|\text{rel}(H, E, C)\rangle$, where rel$(H, C)$ or rel(H,E,C) is the model of the protein secondary structure. E.g. , a hypothetical secondary code such as $HHCCC$ would correspond to the group $G := \left\langle H, C|H^2C^3\right\rangle$ which is called the modular group. Sometimes the group G corresponds (or is close in its structure) to the fundamental group of a three-dimension manifold $\mathcal{M}$ so that we take $\mathcal{M}$ as a candidate manifold of the protein foldings. For the aforementioned example, the candidate manifold would be the trefoil knot complement.

In previous papers, we could describe the (primary) genetic code by using the characters of an appropriate finite group [5, 6]. Now we find, from several protein examples belonging to highly symmetric complexes, that the secondary code has to obey some structural algebraic constraints relying to free group theory. Our generic algebraic building blocks are the hyperbolic (unoriented) 3-manifold of smallest volume known as the Gieseking manifold [7] -when the secondary code only consists of two letters H and C– and the (unoriented) cartographic group $\mathcal{C}_2$ [8, 9] (alias the two-generator free group)–when the secondary code needs the three letters H, E and C. The consistency of the (primary) genetic code and the secondary code is studied under the light of the Kummer surface that we already assumed to play a role in the quaternary structure of protein complexes [6].

In Sect. 2, we provide a few elements about free group theory, finitely generated subgroups of a free group and the fundamental group of a 3-manifold. We single out the mathematical objects that will be useful for our approach of the secondary structures of proteins.

In Sect. 3, we feature a protein example – the histone H3 of drosophila melanogaster– with a short sequence of 136 aminoacids (136 aa) only comprising H and C segments in the secondary pattern. We compare the results obtained from four different models/softwares and how well they fit the cardinality sequence of subgroups of a few candidate 3-manifolds. The Gieseking manifold $m000$ happens to be a good candidate not only in terms of the cardinality sequence but also in terms of the structure of the corresponding subgroups.

In Sect. 4, we pass to more examples of proteins comprising H, E and C patterns. In Sect 4.1, we look at the secondary pattern of myelin P2 in homo sapiens with 133 aa. In Sect. 4.2, we look at the case of the gamma-carbonic anhydrase (247 aa long) within its 3-fold symmetric complex. Then, in

Sect. 4.3, we study the Hfq protein with 74 aa in each arm of the Hfq 6-fold symmetric complex. In both cases, the best theory for modelling the pattern happens to be the cartographic group $\mathcal{C}_2$. In the latter case, the subgroup sequence of $\mathcal{C}_2$ perfectly fits the secondary pattern of Hfq protein predicted by one particular model. In Sect. 4.4, we study the secondary patterns obtained for proteins belonging to 5-fold and 7-fold symmetric complexes. In particular, we provide the comparison of models for the H2A-H2B complex in nucleoplasmin and the acetylcholine receptor (with $n = 5$) and the Lsm1-7 complex (with $n = 7$).

In Sect. 5, we investigate the nucleosome complex which is 8-fold symmetric. Following our previous work in [5, 6], we find that the nucleosome complex allows to define another model of the genetic code preserving quantum information. In addition, one can map the DNA double helix scaffold of the nucleosome complex to the 16 singular points of a Kummer surface.

In Sect. 6, we briefly mention the absolute Galois group over the rationals $\mathbf{G} = \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ as an object worthwhile to be used in the context of protein sequences.

2. Algebraic grometrical models of secondary structures

Let $G = \langle x_1, x_2, \cdots x_l \rangle$ be the free group on l generators. By definition, elements in the group are words u, that are products of elements of G and their inverses modulo a single defining relation $uu^{-1} = e$, with e the identity element. The index $n := |G : G_s|$ of a subgroup G_s in G counts the number of cosets/copies of G_s that fill up G. A right coset with respect to an element $g \in G$ is defined as $G_s g = \{g_s g : g_s \in G_s\}$ so that the set of right cosets partitions G. In other words, every $g \in G$ belongs to just one right coset. Similar statements holds for left cosets. A transversal is an indexed set of (right) coset representatives for G_s in G, and the coset table is a way to express the action of generators x_i and their inverses on them. The algorithm performing this task is the Coxeter-Todd algorithm [10].

Now we pass to finitely presented groups. It is known that every group is a quotient of some free group. One constructs a finitely presented group fp as the quotient of a free group G by the normal subgroup defined by a set of relations rels between the generators x_l

$$fp := \langle x_1, x_2, \cdots x_l | \mathrm{rels}(x_1, x_2, \cdots x_l) \rangle \, .$$

One also needs to define subgroups of finite index in a fp group. A subgroup G_s of the finitely presented group fp is generated by the words specified by a generator list $L_r = L_1 \cdots L_r$ that may contain words or subgroups. In the following, we are interested by the cardinality sequence $\eta_d(fp)$ that counts the number of subgroups of a finite index d up to some maximal index. This sequence allows us to identify a group fp (potentially as the fundamental group of a 3-manifold).

Then, to a pair (fp, G_s) corresponds the permutation group P that organizes the cosets. With the Todd-Coxeter procedure, one can obtain a permutation representation P of the pair from the action of fp on the coset space. In many cases, the finite group P has a geometrical meaning in the sense that it corresponds to a finite geometry [11].

Finally, the group theoretical approach may be related to the theory of 3-manifolds. According to the Poincaré conjecture (now a theorem) every simply connected closed 3-manifold is homeomorphic to the 3-sphere S^3, alias the house of qubits [12]. But one can dress S^3 as a 3-manifold $\mathcal{M}$ that looses the homeomorphism to S^3 following the work of W. Thurston [13]. For instance, the three-dimensional space surrounding the tubular neighborhood of a knot – the knot complement $S^3 \setminus K$ – is a 3-manifold. Among the invariants characterizing a 3-manifold, there is the fundamental group $\pi_1(\mathcal{M})$ which accounts for the first homotopy of $\mathcal{M}$. Finding a 3-manifold $\mathcal{M}$ whose π_1 is the current fp is a way to identify the nature of the object under study.

Below we introduce two algebraic geometric objects playing a role in our description of protein secondary structures. Both objects lack an orientation. The first object is the hyperbolic 3-manifold of the smallest volume [7, 14]. The second one is Grothendieck's cartographic group [8].

2.1. **The Gieseking manifold m000.** This 3-manifold was described by Gieseking in his 1912 thesis. One takes an ideal regular tetrahedron in the 3-dimensional hyperbolic space, that is a tetrahedron with all four vertices on the sphere at infinity and all dihedral angles equal to $\pi/3$. Then one identifies adjacent faces so that the orientation on the edges match [7, Fig. 1]. The resulting hyperbolic manifold has minimal volume among non-compact hyperbolic manifolds. This volume is Gieseking's constant $\int_0^{2\pi/3} \ln(2\cos(x/2))dx = 1.01494160\cdots$. Remarkably, this constant also equals $\zeta_{\mathbb{Q}(i\sqrt{3})}(2)$, which is the Dedekind zeta function at 2 for the field $\mathbb{Q}(i\sqrt{3})$ [14, 15].

The fundamental group for the Gieseking manifold is denoted m000 in SnapPy software [17]. The fundamental group is

$$\pi_1(m000) := \left\langle x, y | x^2y^2 = yx \right\rangle .$$

The cardinality sequence $\eta_d(\pi_1(m000))$ of subgroups of index $d < 15$ of $\pi_1(m000)$ is given in Table 2. The permutation groups organizing the cosets of subgroups of $\pi_1(m000)$ up to index 10 are in Table 1. The identification of submanifolds follows from SnapPy.

In the next section, we find that a model of the secondary structure in histone H3 (PDB 6PWE_1) (obtained with the software PORTER) is the group

$$G := \left\langle C, H | C^{44}H^{12}C^4H^3C^3H^{12}C^8H^{28}C^7H^{10}C^5 \right\rangle .$$

It is shown in Tables 1 and 2 that this model fits perfectly the Gieseking fundamental group at the first 7 places and approximately at the subsequent 3 places. Up to index 7 the permutation groups P are the same. At index 8, all P's related to subgroups of $\pi_1(m000)$ are also those related to subgroups of G, but A_8 and S_8 which are related to subgroups of G are not in subgroups related to $\pi_1(m000)$. There are also a few differences between subgroups of $\pi_1(m000)$ and G at index 9 and 10.

TABLE 1. The d-coverings ($d = 1..10$) of the Gieseking manifold m000. The corresponding 3-manifolds (3-man) are identified thanks to SnapPy. The finite group P organizing the cosets of the index d fundamental group is given. It is shared by almost all subgroups (see lacking P) of the free group associated to the PORTER model of secondary structures of histone H3 (PDB; 6PWE_1). Some extra groups appear in the PORTER model (see extra P).

index	1	2	3	4	5
3-man	m000	K4a1, ooct02_00001	ntet03_00000	m206, otet04_00002 m204, ntet04_00000	m407, ntet05_00007 m405, ncube01_00001
P	(1,1)	(2,1)	(3,1)	(4,1) (12,3)	(5,1) (20,3)
index	6	7	8	9	10
3-man	s961, otet06_00003 x252, ntet06_00004 ntet06_00005	y886, ntet07_00000	t12839, otet06_00007 t12840, otet08_00002 ntet08_00002		
P	(6,2) (12,3) (24,13)	(7,1)	(8,1) (24,3) ×2 (24,13) (96,70), (192,201)	(9,1) (9,1), (648,705)	(10,2) (10,2), (20,3), G_{14400}
lacking P				(72,39)	(320,1635)
extra P			A_8, S_8	(216,53), A_9, S_9	S_{10}, G_{7200}

2.2. **The cartographic group $\mathcal{C}_2$.** The cartographic group is defined as

$$\mathcal{C}_2 := \left\langle x, y, z | x^2 = y^2 = z^2 = (xz)^2 \right\rangle .$$

The terminology comes from Grothendieck's Esquisse d'un programme [8, 9]. It was motivated by the fact that conjugacy classes of transitive subgroups of the oriented subgroup $\mathcal{C}_2^+$ of index 2 of the unoriented group $\mathcal{C}_2$ can be identified to topological maps on connected, oriented surfaces without boundary, while more generally, conjugacy classes of $\mathcal{C}_2$ can be identified with maps on connected surfaces which may or may not be orientable or have a boundary.

The group $\mathcal{C}_2^+$ was investigated by the first author in relation to quantum contextuality in quantum information [11].

In the section below, the group defined from the PORTER model of the secondary structure in protein Hfq (PDB 1HK9) is as follows

$$G := \left\langle C, H, E | C^8 H^{11} C^4 E^6 C^2 E^{10} C E^7 C^3 E^{13} C^9 \right\rangle .$$

It is shown in Table 3 that this group perfectly fits the cartographic group $\mathcal{C}_2$ in terms of the cardinality of subgroups up to the higher index 7 that could be calculated. In addition, the corresponding permutation groups organizing the cosets of subgroups in both the cases of $\mathcal{C}_2$ and G fit as well.

2.3. **Fundamental groups of 3-manifolds.** Hyperbolic 3-manifolds that can be decomposed into regular ideal tetrahedra (up to 25 for the orientable case and up to 21 for the non-orientable case) have been investigated in [16]. Details can be found in SnapPy [17]. In Tables 2 and 3, we collected a few 3-manifolds whose number of subgroups $\eta_d(\pi(\mathcal{M}))$ of index d of their fundamental group $\pi_1(\mathcal{M})$ is close to that of the group arising from the

secondary structure of the protein in question. For example, the figure-of-eight knot $K_0 = K4a1 = 4_1$, which is the subgroup of index 2 in $\pi_1(m000)$, corresponds to the manifold ooct_00001 in SnapPy (see Tables 1 and 2) and $\Sigma_Y = K_0(0,1)$ is the 0-surgery on K_0 [18].

3. Secondary structure with α helices: Drosophila melanogaster histone H3 (PDB 6PWE_1)

Now we show how the theory of the former section may be applied to concrete secondary structures of proteins. One starts with a simple example with two generators (α helices H and coils C). At the next section, we will study a simple example with three generators (α helices H, β sheets E and coils C). Both examples are generic and provide a good credit to our models based on the unoriented hyperbolic manifold m000 and the unoriented cartographic group $\mathcal{C}_2$.

A review of the state of the art in the modelling of secondary structure is given in [2]. It is admitted that there is a limit imposed on the secondary structure prediction due the somewhat arbitrary definition of three states H, E and C. It is true that there exist other fine structures in the secondary protein pattern such as a 3_{10} helix, a π helix and other structures belonging to DSSP (the Dictionary of Protein Secondary Structures). As a result, the assignment inconsistency would limit the highest accuracy based on three states to about 90%. In practice, the best softwares achieve a precision about 80%.

We used the softwares PSIPRED 4.0 [19], PORTER 4.0 [20], PHYRE2 [21] and RAPTORX [22]. We do not enter into the details about the theory of these softwares. Below, we we find that PORTER 4.0 is often well adapted to our goal of identifying an algebraic secondary structure. PORTER 4.0 uses two cascaded bidirectional recurrent neural networks: one for prediction and one for filtering. The method has been trained and benchmarked by cross-validation on a set of many non redundant proteins.

3.1. The primary (linear) structure.

The mRNA sequence for histone H3 of drosophila melanogaster may be found in [23] with the reference NM_001032216.2. It contains 529 base pairings (529 bp). A convenient way to pass from the NCBI format (with line feeds, numbers and blank spaces) to the bare linear sequence is to make use of a software such as Massager [24]. Then, a reading frame such as Expasy [25] allows to extract the candidate proteins.

The $5'3'$ Frame 1 for sequence NM_001032216.2 is as follows

IVFSNVK–T-TLVKPKSE
MARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRP
GTVALREIRRYQKSTELLIRKLPFQRLVREIAQDFKTDLRFQSSAVM
ALQEASEAYLVGLFEDTNLCAIHAKRVTIMPKDIQLARRIRGERA
-ADTALTCR-SASVLYNRSFS

The partial sequence (in bold) beginning at the start codon M and ending at the stop codon '-' is the histone protein H3 with the NCBI reference NP_001027387.1. It can also be found at the protein data base PDB [1] with reference 6PWE_1. The sequence consists of 136 aminoacids (136 aa).

3.2. **The secondary structure.** According to most models, the secondary structure of histone protein H3 only consists of subsections with an α helix H or a coil C.

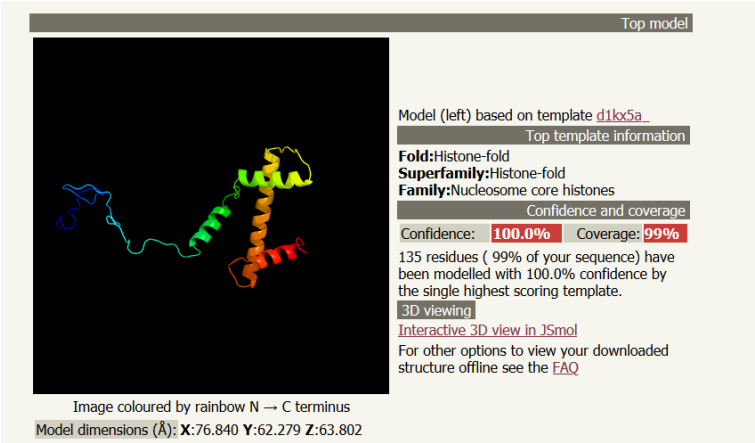

FIGURE 1. A picture of the secondary structure of histone H3 as predicted from PHYRE2.

The predicted secondary structures obtained from the three softwares for the histone H3 protein are as follows

CCCCCCCCCCCCCCCCHHHHCHHHHCCCCCCCCCCCCCCCCCCCCHHHHHHHCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHCCC
CCCCCCCCCCCCCCCCCCCCCHHHHHCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCC

HHHHHCCCCHHHHHHHHHHHCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHCHHHH
CCHHHCCCHHHHHHHHHHHHCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHC
HHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHC
HHHHHHHHHHHHHHHHHHHCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHC

CCCCCCHHHHHHHHHHCCCCC
CCCCCCHHHHHHHHHHCCCCC
CCCCCCHHHHHHHHHHCCCCC
CCCCCCHHHHHHHHHHCCCC

The first line is from PSIPRED, the second one is from PORTER, the third one is from PHYRE2 and the last one is from RAPTORX. One can visually check how close are the predictions.

In table 2, it is found that the best model happens to come from the fundamental group $\pi_1(m000)$ of the Gieseking manifold $m000$ described in subsection 2.1.

4. SECONDARY STRUCTURES WITH α HELICES AND β SHEETS: MYELIN P2, CARBONIC ANHYDRASE AND THE LSM 1-7 COMPLEX

4.1. **Myelin P2 for homo sapiens (PDB 2WUT).** The sequence of myelin P2 in homo sapiens comprises 133 aminoacids as follows. As before, the corresponding four rows for the secondary structures are from PSIPRED, PORTER, PHYRE2 and RAPTORX respectively. One can visually check how close are the predictions.

TABLE 2. The models of the secondary structure for protein H3 of drosophila melanogaster and the cardinality list of d-coverings (alias conjugacy classes of subgroups) of the associated fundamental group. T_1 is the trefoil knot, K_0 is the figure-of-eight knot, the 0-surgery on K_0 is the Akbulut manifold Σ_Y, $\tilde{E}_8$ is the singular fiber of type II* and m000 is the Gieseking manifold. One restricts to two-generator groups since histone H3 only consists of sections with α helices and coils. Observe that the series of cardinalities for the secondary structure of H3 fits the series of the Gieseking manifolds up to the first 7 indices. Bold characters are for partial sequences matching the cardinality sequence of Gieseking manifold m000.

protein	model	$\eta_d(T)$
H3 (6PWE_1)	PSIPRED	[1,1,1,1,2, 2,1,3,5,5 .,.,.,.,.]
H3	PHYRE2	[1,1,1,1,3, 4,1,5,10,10 .,.,.,.,.]
H3	PORTER	[**1,1,1,2,2, 3,1**,12,6,5 .,.,.,.,.]
H3	RAPTORX	[1,1,1,1,2, 1,1,2,3,3 .,.,.,.,.]
m000	Gieseking	[**1,1,1,2,2, 3,1**,4,3,5, 4,14,1,5,10]
T_1	trefoil	[1,1,2,3,2, 8,7,10,18,28, 27,88,134,171,354]
K_0	figure-of-eight	[**1,1,1,2**,4, 11,9,10,11,38, 26,62,39,89,228]
$K_0(0,1)$	Σ_Y	[**1,1,1,2,2** ,5,1,2,2,4, 3,17,1,1,2]
$\tilde{E}_8$	singular fiber II*	[1,1,2,2,1 ,5,3,2,4,1, 1,12,3,3,4]

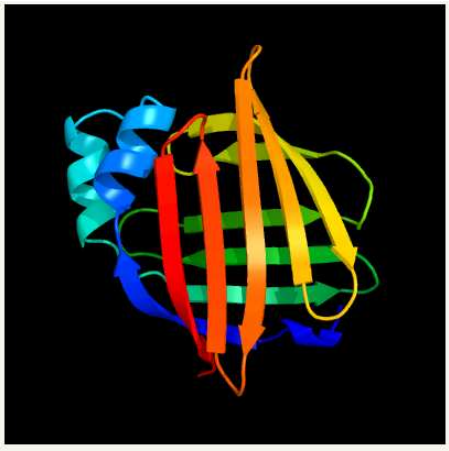

FIGURE 2. A picture of the secondary structure of myelin P2 in homo sapiens (PDB 2WUT) as predicted from PHYRE2.

GMSNKFLGTWKLVSSENFDDYMKALGVGLATRKLGNLAKPTVIISKKGDIITIRTESTFKN
CCCHHCCEEEEEEEECCHHHHHHHCCCCHHHHHHHHHCCCEEEEEEECCEEEEEEECCCC
CCCHHCCEEEEEECCCCHHHHHHHCCCCHHHHHHHHHCCCEEEEEEECCEEEEEEECCCC
CCCCCCEEEEEEEEECCHHHHHHHHCCCHHHHHHHHCCCCEEEEEEECCEEEEEEECCCC
CCCCCCEEEEEEEEECCHHHHHHHCCCCHHHHHHHHCCCCEEEEEEECCEEEEEEECCCC

TEISFKLGQEFEETTADNRKTKSIVTLQRGSLNQVQRWDGKETTIKRKLVNGKMVAECKM
CCCHHCCEEEEEEEECCHHHHHHHCCCCHHHHHHHHHCCCEEEEEEECCEEEEEEECCCC

EEEEEEEECCEEEEECCCCCEEEEEEEEECCEEEEEECCCCCEEEEEEEEECCEEEEEEE
EEEEEEECCCEEEEECCCCCEEEEEEEEECCEEEEEECCCCCEEEEEEEECCEEEEEEE
EEEEEEECCCEEEEECCCCCEEEEEEEEECCEEEEEECCCCCEEEEEEEECCEEEEEEE

KGVVCTRIYEKV
CCEEEEEEEEEC
CCEEEEEEEEEC
CCEEEEEEEEEC
CCEEEEEEEEEC

Using Table 3, one observes that the cardinality sequence of subgroups in the PHYRE2 and PORTER models of the secondary structure of myelin P2 corresponds to that of the cartographic group C_2 up to index 4. Up to this index, one can also show that the permutation groups P for the structure of cosets in PHYRE2 and PORTER models correspond to that of $\mathcal{C}_2$.

4.2. **The 3-fold symmetric complex for gamma-carbonic anhydrase (PDB 1QRE).** In the protein data bank the gamma-carbonic anhydrase for methanosarcina thermophila (PDB 1QRE_1) is a sequence with 247 aa. As for myelin P2, using Table 3, one observes that the cardinality sequence of subgroups in the PHYRE2 and PORTER models of the secondary structure of 1QRE_1 corresponds to that of the cartographic group C_2 up to index 4. The complex is 3-fold symmetric as shown in Fig. 3a.

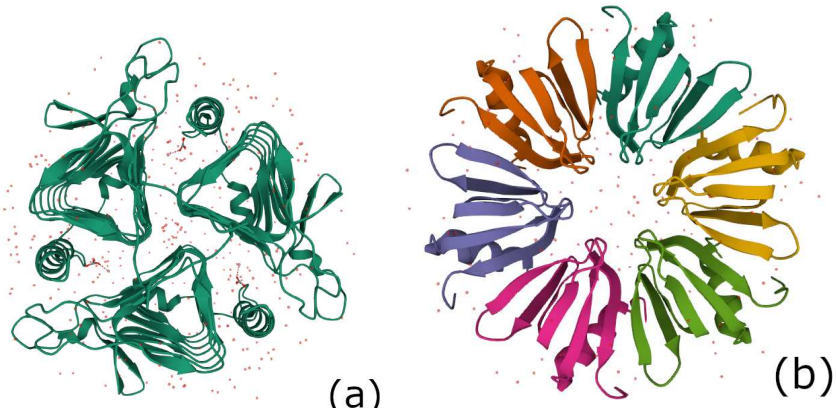

FIGURE 3. (a) A picture of the structure of carbonic anhydrase (PDB 1QRE), (b) A picture of the structure of Hfq protein complex of Escherichia coli (PDB 1HK9) .

4.3. **The Hfq protein complex of Escherichia coli (PDB 1HK9).** The sequence of Hfq protein of Escherichia coli (PDB 1HK9_1) comprises 74 aminoacids. As before, the corresponding four rows for the secondary structures are from PSIPRED, PORTER, PHYRE2 and RAPTORX respectively. One can visually check how close are the predictions.

GAMAKGQSLQDPFLNALRRERVPVSIYLVNGIKLQGQIESFDQFVILLKNTVSQMVYKHAISTVVPSRPVSHHS
CCCCCCCCCCHHHHHHHHHHCCCCEEEEEECCCEEEEEEEECCCEEEEEECCCEEEEEEEEEEEEECCCCCCCCC
CCCCCCCCHHHHHHHHHHHCCCCEEEEEECCEEEEEEEEEECEEEEEEECCCEEEEEEEEEEEEECCCCCCCCC
CCCCCCCCCHHHHHHHHHHCCCEEEEEECCEEEEEEEEEECCEEEEEECCCCEEEEEEEEEEEEECCEEEECC
CCCCCCCCCCHHHHHHHHHCCCCEEEEECCCCEEEEEEEEECCCEEEEEECCCEEEEEEEEEEEEECCCCCCCC

The PORTER model for this protein happens to coincide with that of the cartographic group $\mathcal{C}_2$ described in the subsection 2.2.

As shown in Fig. 3b, the Hfq complex consists of a quaternary structure with 6-fold symmetry where each arm contains the protein Hfq. This object was studied in our recent paper [6, Sect. 2.2] as leading to a Kummer surface related to the character table of the finite group $G_6 = (288, 69) \equiv \mathbb{Z}_6 \rtimes (2O)$.

4.4. **Other n-fold symmetric complexes.**

The 5*-fold symmetric H2A-H2B complex in nucleoplasmin (PDB 2XQL).* Molecular chaperones are proteins that help the folding or unfolding and the disassembly of other molecular structures. Nucleoplasmin, the first identified molecular chaperone, promotes the in vitro assembly of nucleosomes. The latter are the topic of our next section. There is a histone octamer comprising two H2A-H2B dimers and an H3-H4 tetramer. The H2A-H2B histone complex is investigated in [26]. It has a pentameric structure as shown in Fig. 4a and is referred as 2XQL in the protein data bank.

We performed an investigation of the secondary structure of the 2XQL_1 protein that one finds in each of the 5 arms of the complex. PSIPRED and PORTER models predict a secondary structure with α helices and coils only that we could not compare to a known group theoretical sequence. The PHYRE2 and RAPTORX models, as well as our approach based on the mapping of aminoacids to the characters of group G_7 and G_8 (explained below), predict a cardinality sequence which fits that of the cartographic group $\mathcal{C}_2$, as shown in Table 3.

The 5*-fold symmetric acetylcholine receptor (PDB 2BG9).* The acetylcholine receptor is an integral membrane protein that responds to the binding of the acetylcholine neurotransmitter. This receptor is also sensitive to nicotine and muscarine. It has a pentameric structure shown in Fig. 4b and is refereed as 2BG9 in the protein data bank.

We performed an investigation of the secondary structure of the 2BG9_1 protein that one finds in the 5 arms of the complex. As shown in Table 3, all models predict a secondary structure with α helices, β sheets and coils. One does not observe a good fit to a group theoretical structure shared by all models. The best fit is between the RAPTORX model and the fundamental group of the 3-manifold ooct_00001 where the cardinality (and the structure) of subgroups coincide up to 4 places.

The 7*-fold symmetric Lsm 1-7 complex in the spliceosome (PDB 4M75).* In molecular biology, there exists an ubiquitous family of RNA-binding proteins called LSM proteins whose function is to serve as scaffolds for RNA oligonucleotides, assisting the RNA to maintain the proper three-dimensional structure. Such proteins organize as rings of six or seven subunits. The Hfq protein complex was discovered in 1968 as an Escherichia coli host factor that was essential for replication of the bacteriophage $Q\beta$ [27], it displays an hexameric ring shape shown in Fig. 3b of the previous subsection. As already mentioned it is remarkable that the secondary structure of Hfq protein is so close to the cartographic group model.

TABLE 3. A few proteins, the model of their secondary structure and the cardinality list of d-coverings (alias conjugacy classes of subgroups of index d) of the associated fundamental group. One takes proteins that contain sections with α helices, β sheets and coils. The groups obtained by mapping the appropriate characters of $G_7 = (336, 118)$ and $G_8 = (384, 5589)$ to amino acids are also considered. Bold characters are for partial sequences matching the sequence of the cartographic group $\mathcal{C}_2$.

protein	aa	model	$\eta_d(T)$
myelin P2 (2WUT)	133	PSIPRED	[1,3,13,84,336, 4216]
2WUT		PHYRE2	[**1,3,7,26**,164, 10669]
2WUT		PORTER	[**1,3,7,26**,135, 871]
2WUT		RAPTORX	[1,3,10,59,348, 2899]
.		(336,118)	[1,3,7,30,122, 991]
.		(384,5589)	[1,3,7,34,130, 999]
carbonic anhydrase (1QRE_1)	247	PSIPRED	[1,3,10,43,135, 1071]
1QRE_1		PHYRE2	[**1,3,7,26**,149, 1085]
1QRE_1		PORTER	[**1,3,7,26**,415, 4382]
1QRE_1		RAPTORX	[1,3,10,35,106, 804]
.		(336,118)	[1,3,7,30,150, 883]
.		(384,5589)	[1,3,10,47,148, 1015]
protein Hfq (1HK9_1)	74	PSIPRED	[1,7,17,114,1145 14275]
1HK9_1		PHYRE2	[1,7,14,149,1458, 21756]
1HK9_1		PORTER	[**1,3,7,26,97, 624,4163,34470**]
1HK9_1		RAPTORX	[1,3,10,51,162, 1434]
.		(336,118)	[**1,3,7,26**,134, 912]
.		(384,5589)	[1,3,7,34,146 894]
H2A-H2B (2XQL_1)	91	PHYRE2	[**1,3,7,26**,103, 688]
2XQL_1		RAPTORX	[**1,3,7,26**,165, 2272]
.		(336,118)	[**1,3,7,26**,130, 943]
.		(384,5589)	[**1,3,7,26**,136, 967]
acetylcholin receptor (2BG9_1)	370	PSIPRED	[1,3,10,35,151, 1023]
2BG9_1		PHYRE2	[1,7,11,92,288, 2087]
2BG9_1		PORTER	[1,7,11,92,239, 2058]
2BG9_1		RAPTORX	[1,3,7,34,169, 1432]
.		(336, 118)	[1,3,10,47,124, 1026]
.		(384, 5589)	[1,3,7,30,140, 931]
Lsm 1-7 complex (4M75_1)	144	PSIPRED	[1,3,16,81,184, 1800]
4M75_1		PHYRE2	[1,7,14,201,705, 8850]
4M75_1		PORTER	[**1,3,7,26**,139, 1118]
4M75_1		RAPTORX	[**1,3,7,26**,125, 747]
.		(336, 118)	[1,3,7,34,145, 948]
.		(384, 5589)	[1,3,10,35,135, 975]
$\mathcal{C}_2$	na	cartographic group	[**1,3,7,26,97, 624, 4163, 34470**]
ooct02_00017		3-manifold	[**1,3,7,26**,40, 231]
ooct02_00006		3-manifold	[1,3,10,43,112, 802]
noct02_00024		3-manifold	[1,3,10,43,117, 804]
ooct02_00009		3-manifold	[1,3,7,30,105, 649]
ooct04_00001		3-manifold	[1,3,7,34,43,240, 254]
L7a1		3-manifold link	[1,3,7,34,75,377, 807]
ooct03_00019		3-manifold	[1,7,11,85,95,240, 492]

It is known that, in the process of transcription of DNA to proteins through messenger RNA sequences (mRNAs), there is an important step

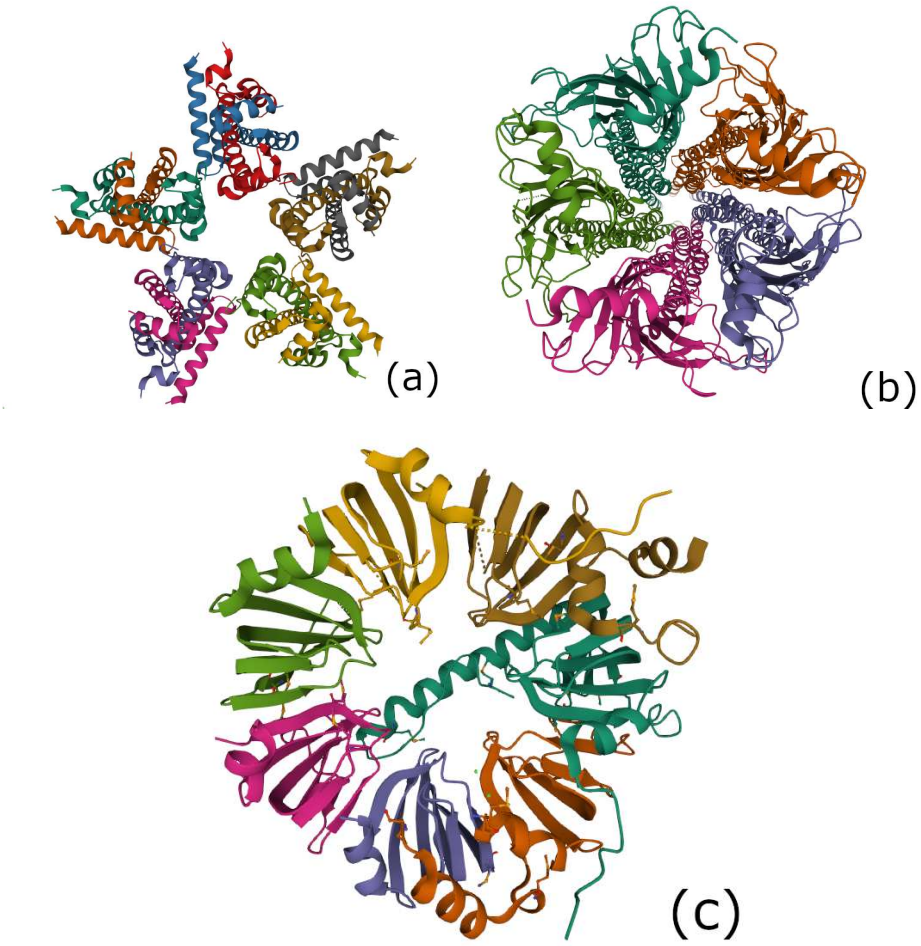

FIGURE 4. (a) the nucleoplasmin H2A-H2B: 2XQL in the protein data bank, (b) the acetylcholine receptor: 2BG9 in the protein data bank, (c) the Lsm 1-7 complex in the spliceosome: 4M75 in the protein data bank

performed in the spliceosome [28]. It includes removing the non-coding intron sequences for obtaining the exons that code for the proteinogenic amino acids. A ribonucleoprotein (RNP) –a complex of ribonucleic acid and RNA-binding protein– plays a vital role in a number of biological functions that include transcription, translation, the regulation of gene expression and the metabolism of RNA. Individual LSm proteins assemble into a six or seven member doughnut ring which usually binds to a small RNA molecule to form a ribonucleoprotein complex.

In our previous paper [6], it was shown that 7-fold symmetry may be mirrored in the finite group $G_7 = \mathbb{Z}_7 \times 2O$ (with $2O$ the binary octahedral group) whose characters may be mapped to the amino acids of the genetic code. Such a mapping is reproduced in Table 4. It is important to mention that the characters of G_7 are informationally complete except for the trivial

character that is not used in the mapping to aminoacids and the character mapped to the starting amino acid **M**.

It was also determined an algebraic object called a Kummer surface playing a role in the mapping of characters to aminoacids.

(336,118)	dimension	1	1	1	2	2	2	2	2	2	2
$\mathbb{Z}_7 \rtimes (\mathbb{Z}_2.S_4)$	d-dit, d=29	29	785	d^2	d^2	d^2	d^2	d^2	d^2	d^2	d^2
$\cong \mathbb{Z}_7 \rtimes 2O$	amino acid	.	**M**	**W**	**C**	**F**	**Y**	.	.	**H**	**Q**
	order	1	2	3	4	4	6	7	7	7	8
	char	Cte	Cte	Cte	z_1	z_1	z_1	z_4	z_4	$z_{1,5}$	$z_{1,5}$
	polar req.	.	5.3	5.2	4.8	5.0	5.4	.	.	8.4	8.6
(336,118)	dimension	2	2	2	2	3	3	4	4	4	4
	d-dit, d=29	d^2	d^2	d^2	d^2	d^2	d^2	d^2	d^2	d^2	d^2
	amino acid	**N**	**K**	**E**	**D**	**I**	Stop	.	.	.	.
	order	14	14	14	21	21	21	21	21	21	21
	char	$z_{1,5}$	$z_{1,5}$	$z_{1,5}$	$z_{1,5}$	Cte	Cte	Cte	$z_{1,2}$	$z_{1,2}$	$z_{1,2}$
	polar req.	10.0	10.1	12.5	13.0	10	15	.	.	.	.
(336,118)	dimension	4	4	4	f 4	4	4	6	6	6	
	d-dit, d=29	d^2	d^2	d^2	d^2	d^2	d^2	d^2	d^2	d^2	
	amino acid	**V**	**P**	**T**	**A**	**G**	.	**L**	**S**	**R**	
	order	28	28	28	42	42	42	42	42	42	
	char	$z_{2,5}$	$z_{2,5}$	$z_{2,5}$	$z_{2,5}$	$z_{2,5}$	$z_{2,5}$	$z_{1,3}$	$z_{1,3}$	$z_{1,3}$	
	polar req.	5.6	6.6	6.6	7.0	7.9	.	4.9	7.5	9.1	

TABLE 4. For the group $G_7 := (336, 118) \cong \mathbb{Z}_7 \rtimes 2O$, the table provides the dimension of the representation, the rank of the Gram matrix obtained under the action of the 29-dimensional Pauli group, the order of a group element in the class, the angles involved in the character and a good assignment to an amino acid according to its polar requirement value. All characters are informationally complete except for the trivial character and the one assigned to **M**. The entries involved in the characters are $z_1 = 2\cos(2\pi/7)$, $z_2 = 2z_1$, $z_3 = -6\cos(\pi/7)$, $z_4 = \sqrt{2}$ and $z_5 = 2\cos(2\pi/21)$ featuring the angles $2\pi/8$ (in z_4), $2\pi/7$ and $2\pi/21$.

Encoding a protein with the characters of the finite field G_7. Since the group G_7 is successful for encoding the genetic code and that, at the same time, it provides an assignment to the 20 amino acids through the corresponding characters, one can ask ourselves if G_7 may also be used to define a secondary structure in a protein. Indeed we can get a secondary structure from the character table in the following way.

Observe that, to a character in Table 4, corresponds an entry denoted z_1, z_4, $z_{1,2}$, $z_{1,3}$, $z_{1,5}$ or $z_{2,5}$ which expresses which z_i appears in the slot/character. This entry mainly reflects the character field associated to the character. For example, there are 11 slots (and 11 amino acids) containing z_5 and from these characters one can also define the aforementioned Kummer surface. Let us choose to assign to these slots a secondary structure H_0 and to assign a secondary structure C_0 to the remaining slots encoding an amino acid. This method allows to encode the protein under examination with pseudo-helices H_0 and pseudo-coils C_0.

We can refine the technique by introducing more structure in the pseudo coil segments. Some of the slots/amino acids correspond to a character with

constant entries and we choose to encode them as C_0 as before and the remaining slots/amino acids which correspond to a non constant entry (z_1 or $z_{1,3}$) are encoded with E_0, that we consider as a pseudo-sheet.

Then we can define the group

$G_0 := \langle H_0, E_0, C_0 | \mathrm{rel}(H_0, E_0, C_0) \rangle$, where $\mathrm{rel}(H_0, E_0, C_0)$ is the new model of the protein secondary structure obtained by our definition of pseudo-helices H_0, pseudo-sheets E_0 and pseudo-coils C_0. In table 3, the cardinality structure of group G_0 is compared to that of the other models PSIPRED, PHYRE2, PORTER and RAPTORX. One finds that the cardinality sequence either fits, at the first few places, the cartographic group $\mathcal{C}_2$ or that of a 3-manifold. It leaves open the question whether one of the standard models or our own model is the most efficient.

5. The 8-fold symmetric histone complex of the nucleosome: 3WKJ in the protein data bank

Strong DNA packaging is found in the nucleosome of eukaryotes. The nucleosome complex consists of a double helix wrapped around a set of eight histone proteins comprising two copies of H2A, H2B, H3 and H4. The nucleosome is the fundamental subunit of chromatin. Eukaryotic chromatin is further compacted by being folded into more complex structures eventually forming a chromosome. Nucleosomes are considered to be the support of epigenetic information. The nucleosome core particle contains approximately 146 base pairs (bp) of DNA wrapped in 1.67 left-handed superhelical turns around the histone octamer as shown in Fig. 5a.

We already met histone H3 of a different specie (drosophila melanogaster) in Sect. 3 as the preliminary example of a protein only containing α helices and random coils. In the histone complex 3WKJ of the nucleosome, the secondary structure of histone H3 is also found to be made of segments with α helices and coils but with a different organization according to our group theoretical approach. This is also true for the other histones H4, H2A and H2B of the histone octamer.

In this section, we do not enter into the secondary structure of histones. We rather focus on the 8-fold symmetry of the core particle in the histone complex. What interests us about the double helix is the fact that their projection is a set of 16 double points as shown by the arrows in Fig. 5a. The reader may be familiar with our previous paper [6] in which 16 double points occur in a beautiful algebraic object called a Kummer surface. Such a Kummer surface was constructed from the character table of the group $G_7 = (336, 118) \cong \mathbb{Z}_7 \rtimes 2O$ in the context of the spliceosome complex that we investigated in Sect. 4.4. Below, we pursue in the same line of ideas and build another model of the genetic code based on the group $G_8 = (384, 5589) \cong \mathbb{Z}_8 \rtimes 2O$ and a corresponding Kummer surface.

The character table for the group G_8 is in Table 5. As before for the group G_7, Table 5 contains a good assignment to the 20 amino acids and some details about the character fields through the entries z_i. For dimensions 2 and 4, the assignments correspond to characters that are informationally complete. But it is not the case for the assigments of amino acids in dimensions 1, 3 and 6.

(384,5589)	dimension	1	1	1	1	2	2	2	2	2	2
$\mathbb{Z}_8 \rtimes (48, 28)$	d-dit, d=37	37	1333	1333	1333	1361	d^2	d^2	1367	d^2	d^2.
$\cong \mathbb{Z}_8 \rtimes 2O$	amino acid	.	.	**M**	**W**	.	.	.	.	.	.
	char	Cte	Cte	Cte	Cte	Cte	Cte	Cte	z_1	z_1	z_1
(384,5589)	dimension	2	2	2	2	2	2	2	2	2	3
	d-dit, d=37	d^2	d^2	d^2	d^2	d^2	d^2	d^2	d^2	d^2	1367
	amino acid	**C**	**F**	**Y**	**H**	**Q**	**N**	**K**	**E**	**D**	.
	char	z_1	z_1	z_1	z_4	z_4	$z_{1,4,5}$	$z_{1,4,5}$	$z_{1,4,5}$	$z_{1,4,5}$	Cte
(384,5589)	dimension	3	3	3	4	4	4	4	4	4	4
	d-dit, d=37	d^2	1367	1367	d^2	1367	1367	d^2	d^2	d^2	d^2
	amino acid **I**	Stop	.	.	.	.	.	.	.	**V**	
	char	Cte	Cte	Cte	Cte	Cte	Cte	$z_{1,2}$	$z_{1,2}$	z_4	z_4
(384,5589)	dimension	4	4	4	4	6	6	6			
	d-dit, d=37	d^2	d^2	d^2	d^2	701	1365	1365			
	amino acid	**P**	**T**	**A**	**G**	**L**	**S**	**R**			
	char	$z_{2,4,5}$	$z_{2,4,5}$	$z_{2,4,5}$	$z_{2,4,5}$	Cte	$z_{1,3}$	$z_{1,3}$			

TABLE 5. For the group $G = (384, 5589) \cong \mathbb{Z}_8 \rtimes 2O$, the table provides the dimension of the representation, the rank of the Gram matrix obtained under the action of the 37-dimensional Pauli group and the entries involved in the characters. The notation is $z_1 = -\sqrt{2}$, $z_2 = 2\sqrt{2}$, $z_3 = 3\sqrt{2}$, $z_4 = -\sqrt{3}$ and $z_5 = -2\cos(5\pi/12)$. All characters having z_4 and z_5 in their entries are informationally complete and are at the origin of the Kummer surface. All characters having entries with z_2 or z_4 are also informationally complete. A good matching to the aminoacids (ordered according to their polar requirement and simultaneously to the order of a group element) is given.

All 8 characters having $z_4 = \sqrt{3}$ and $z_5 = -2\cos(5\pi/12)$ in their entries are informationally complete and are at the origin of the Kummer surface. We now show an important characteristics of such characters. As an example, let us write the character number 16 as obtained from Magma [10]

$$\kappa_{16} = [2, -2, -2, 2, -1, 0, 0, 2, -2, 0, 0, 0, 1, -1, 1, z_1, -z_1, z_1, -z_1, z_1, -z_1$$
$$0, 0, 0, 0, z_4, -z_4, -z_4, z_4, z_5, z_5, z_5\#5, z_5\#5, -z_5, -z_5\#5, -z_5 - z_5\#5]$$

where # denotes the algebraic conjugation, that is $\#k$ indicates replacing the root of unity w by w^k.

One defines a genus 2 hyperelliptic curve $\mathcal{C}_8 : y^2 = f(x)$ defined over the group G_8 from the equation

$$y^2 = f(x) = (x + k)(x - k)(x + l)(x - l)(x + m)(x - m),$$

with $k = \sqrt{3}$, $l = 2\cos(5\pi/12)$ and $m = 2\cos(\pi/12)$. Explicitely,

$$\mathcal{C}_8 : \; y^2 = x^6 - 7x^4 + 13x^2 - 3,$$

leading to the polynomial definition of the Kummer surface $S(x_1, x_2, x_3, x_4)$ as

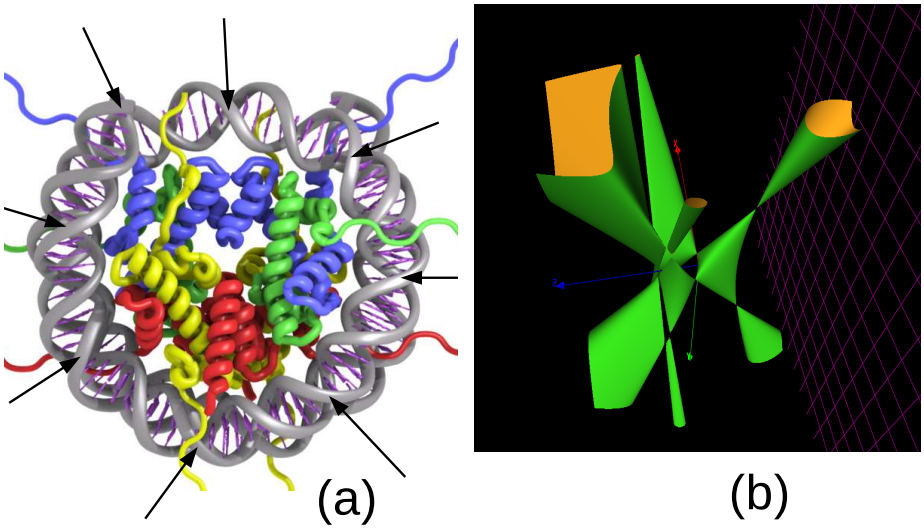

FIGURE 5. (a) The structure of a nucleosome consists of a DNA double helix wound around eight histone proteins. There are eight periods (as shown in the picture) so that the two helices meet at 16 points . They map to the 16 double points of the Kummer surface. (b) A section at constant x_4 of the Kummer surface for the group G_8.

$$S(x_1,x_2,x_3,x_4) = 156x_1^4 + 12x_1^3x_4 - 84x_1^2x_2^2 + 376x_1^2x_3^2 - 52x_1^2x_3x_4$$
$$24x_1x_2^2x_3 + 28x_1x_3^2x_4 - 4x_1x_3x_4^2 + 12x_2^4 - 52x_2^2x_3^2 + x_2^2x_4^2 + 28x_3^4 - 4x_3^3x_4.$$

The desingularisation of the Kummer surface is obtained in a simple way by restricting the product $f(x)$ to the five first factors.

As usual for elliptic and hyperelliptic curves of genus g, $\mathcal{C}_8$ is embedded in a weighted projective plane, with weights 1, $g+1$, and 1, respectively on coordinates x, y and z. Therefore, point triples are such that $(x:y:z) = (\mu x:\mu y:\mu z)$, μ in the field of definition, and the points at infinity take the form $(1:y:0)$. Below, the software Magma is used for the calculation of points of $\mathcal{C}_8$ [10]. For the points of $\mathcal{C}_8$, there is a parameter called 'bound' that loosely follows the heights of the x-coordinates found by the search algorithm.

It is found that the corresponding Jacobian of $\mathcal{C}_8$ has $16 = 6 + 10$ points as follows

* the 6 points bounded by the modulus 1:

$Id := (1,0,0)$, $K_{\pm 1} := (x \pm k, 0, 1)$, $L_{\pm 1} := (x \pm l, 0, 1)$ and $M = (x - m, 0, 1)$.

* the 10 points of modulus > 1:

$a_1 := K_1 + K_{-1}$, $a_2 := K_1 + M$, $a_3 := K_1 + K_{-1} + L_1$, $a_4 := K_1 + L_1$, $a_5 := K_{-1} + M$, $a_6 := K_1 + K_{-1} + L_{-1}$, $a_7 := K_{-1} + L_1$, $a_8 := K_{-1} + L_{-1}$, $a_9 := K_1 + K_{-1} + M$ and $a_{10} := K_1 + L_{-1}$.

The 16 points organize as a commutative group isomorphic to the maximally abelian group $\mathbb{Z}_2^4$ as shown in the following Jacobian addition table

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

TABLE 6. The structure of the addition table for the 16 singular Jacobian points of the hyperelliptic curves $\mathcal{C}_8$.

where the blocks are given explicitely as

$$A : \begin{bmatrix} Id & K_1 & K_{-1} & a_1 \\ K_1 & Id & a_1 & K_{-1} \\ K_{-1} & a_1 & Id & K_1 \\ a_1 & K_{-1} & K_1 & Id \end{bmatrix}, \; B : \begin{bmatrix} M & a_2 & a_5 & a_9 \\ a_2 & M & a_9 & a_5 \\ a_5 & a_9 & M & a_2 \\ a_9 & a_5 & a_2 & M \end{bmatrix},$$

$$C : \begin{bmatrix} L_1 & a_4 & a_7 & a_3 \\ a_4 & L_1 & a_3 & a_7 \\ a_7 & a_3 & L_1 & a_4 \\ a_3 & a_7 & a_4 & L_1 \end{bmatrix}, \; D : \begin{bmatrix} a_6 & a_8 & a_{10} & L_{-1} \\ a_8 & a_6 & L_{-1} & a_{10} \\ a_{10} & L_{-1} & a_6 & a_8 \\ L_{-1} & a_{10} & a_8 & a_6 \end{bmatrix}.$$

To conclude this section, we can define a model of the secondary structure of nucleosome complex based on the character table of G_8 as we did for the spliceosome complex with the character table of G_7. The amino acids that are mapped to characters containing z_5 should belong to a pseudo-helix H_0 of the secondary structure. The other amino acids either correspond to a constant entry in the character table and belong to a pseudo-coil C_0 or to a non-constant entry (which is either z_1, z_4 or $z_{1,3}$) and belong to a pseudo-sheet E_0. In table 3, the cardinality structure of group G_0 obtained with this model is compared to that of the other models PSIPRED, PHYRE2, PORTER and RAPTORX. One again observes that the cardinality sequence either fits, at the first few places, the cartographic group $\mathcal{C}_2$ or that of a 3-manifold.

6. Discussion

The (primary) genetic code maps the 4-base words of DNA to the 20 proteinogenic amino acids, a feature that we could model by using concepts of quantum information theory associated to finite group representations. The (mostly informationally complete) characters of finite groups G_n of signature $\mathbb{Z}_n \rtimes 2O$ ($2O$ the binary octahedral group) are able to account for the degeneracies and many properties of the code

(see [5] when $n = 5$, [6] when $n = 6$ and Sect. 5 of this paper when $n = 7$).

The secondary 'genetic code' lacks the universality of the primary code. In the standard models of the secondary structure of proteins, the mapping from the 20 amino acids to segments of α helices H, β sheet strands E and coils C is not pointwise. The present generation of softwares is defined by the evolutionary information derived from alignment of multiple homologous sequences and the highest reported accuracy uses neural networks for the optimal comparison of the sequences [2].

We could identify algebraic structures in the secondary code by employing the theory of infinite groups with generators H, E and C and the protein relation induced by the chosen model. It was unexpected that the cartographic group $\mathcal{C}_2$ seems to play a major role in the secondary structure. Why are we interested by this feature?

We are interested in geometric physical codes or languages in action [29] and their connection to the concept of emergence. Group representations arise here as a formal way to describe those geometrical codes. Back to the cartographic group, we already mentioned in the introduction that maps on surfaces which may be non-orientable or with boundary correspond to $\mathcal{C}_2$. Another important aspect is that $\mathcal{C}_2$ is related to the so called absolute Galois group $\mathbf{G} = \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$, the group of field-automorphisms of the field extension $\bar{\mathbb{Q}}$ of the rational field $\mathbb{Q}$. In the *Esquisse d'un programme* [8, 9, 30], Grothendieck emphasizes the interest of looking at the action of $\mathbf{G}$ on topological, geometric and even combinatorial structures. The highest level is the so-called 'Teichmüller tower'. The simplest level concerns bipartite (hyper)maps called 'dessins d'enfants'. To any dessin $\mathcal{D}$ corresponds a (so-called) Belyi function $f(x)$, where $f(x)$ is a rational function of the complex variable x whose structure reflects the critical points and the topology of $\mathcal{D}$. The remarkable result is that $\mathbf{G}$ acts faithfully on $\mathcal{D}$, that is, each non-identity element of $\mathbf{G}$ sends two non-isomorphic dessins to two inequivalent Belyi functions $f(x)$, so that none of the structure of $\mathbf{G}$ is lost by proceeding in this way. In passing, it is good to mention that the theory of 'dessins d'enfants' can be used to account for geometry contextuality, the counterpart of quantum contextuality [11, 31].

Let us go back to the secondary structure of protein Hfq in Sect. 4.3 that builds one of the 7 arms of the Lsm1-7 complex in Fig. 3b. According to our theory, there is a group structure of the protein that intimately reflects that of $\mathcal{C}_2$. Every subgroup of index d of $\mathcal{C}_2$ can be seen as permutation group on d elements, it can be drawn as a dessin $\mathcal{D}$ and there is a faithful action of $\mathbf{G}$ on all dessins/permutation groups. In other words, the protein Hfq contains in its structure the topology and algebra of $\mathbf{G}$. The biological meaning of this algebraic geometric structure needs further work. We leave it open at this stage. It may

be that the constraint of approximating the secondary structure with three letter segments H, E and C implies that every protein has to obey the **G** rules. We believe that this rule may be seen as a support of the connection of biology to quantum gravity. In [33], it is shown how a theory of quantum gravity may connect to **G**. We already proposed a connection of our approach of the genetic code (see [6] and Sect. 5 of this paper) to the Kummer surfaces that are K_3 surfaces and play a role in some models of quantum gravity [32].

References

[1] The protein data bank, available at https://pdb101.rcsb.org/

[2] Y. Dang, J. Gao, J. Wang, R. Heffernan, J. Hanson, K. Paliwal and Y. Zhou, Sixty-five years of the long march in protein secondary structure prediction: the final strech? *Briefings in Bioinformatics* **19** (3) 482–494 (2018).

[3] L. Pauling, R. B. Corey and H. R. Branson, The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain, *Proc. Natl. Acad. Sci.* **37** 205–211 (1951).

[4] L. Pauling and R. B. Corey, Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets, *Proc. Nat. Acad. Sci.* **37** 729–740 (1951).

[5] M. Planat, R. Aschheim, M. M. Amaral, F. Fang and K. Irwin, Complete quantum information in the DNA genetic code, *Symmetry* **12** 1993 (2020).

[6] M. Planat, D. Chester, R. Aschheim, M. M. Amaral, F. Fang and K. Irwin, Finite groups for the Kummer surface: the genetic code and quantum gravity, *Quantum Reports* **3** 68–79 (2021).

[7] C. C. Adams, The noncompact hyperbolic 3-manifold of minimal volume, *Proc. Am. Math. Soc.* **4** 100 (1987).

[8] A. Grothendieck, Sketch of a programme, written in 1984 and reprinted with translation in L. Schneps ans P. Lochak eds, Geometric Galois Actions 1. Around Grothendieck's Esquissed'un Programme, 2. The inverse Galois problem, Moduli Spaces and Mapping Class Groups(Cambridge University Press, 1997); (b) The Grothendieck Theory of Dessins d'Enfants, L.Schneps (ed) (Cambridge Univ. Press, 1994).

[9] S. K. Lando and A. K. Zvonkin, *Graphs on surfaces and their applications* (Springer Verlag, Berlin, 2004).

[10] Bosma, W.; Cannon, J. J.; Fieker, C. ; Steel, A. (eds). *Handbook of Magma functions*, Edition 2.23 (2017), 5914pp (accessed on 1 January 2021).

[11] M. Planat, A. Giorgetti, F. Holweck and M. Saniga, Quantum contextual finite geometries from dessins d'enfants, *Int. J. Geom. Mod. Phys.* **12** 1550067 (2015).

[12] M. Planat, R. Aschheim, M. M. Amaral and K. Irwin, Universal quantum computing and three-manifolds, Universal quantum computing and three-manifolds *Symmetry* **10** 773 (2018).

[13] W. P. Thurston, *Three-dimensional geometry and topology*, (Princeton University Press, Princeton, N.J., USA, 1997), vol. 1.

[14] C. C. Adams, The newest inductee in the number hall of fame, *Math. Mag.* **71** 341–349 (1998).

[15] J. Milnor, Hyperbolic geometry: the first 150 years, *Bull. AMS* **6** 9–24 (1982).

[16] E. Fominikh, S. Garoufalidis, M. Goerner, V. Tarkaev and A. Vesnin, A census of tetrahedral hyperbolic manifolds, *Exp. Math.* **25** 466-481 (2016).

[17] M. Culler, N. M. Dunfield, M. Goerner, and J. R. Weeks, SnapPy, a computer program for studying the geometry and topology of 3-manifolds, http://snappy.math.uic.edu/ (accessed on 1 January 2021).

[18] M. Planat, R. Aschheim, M. M. Amaral and K. Irwin, Quantum computing, Seifert surfaces and singular fibers, *Quantum Reports* **1** 12-22 (2019).

[19] D T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, *J. Mol. Biol.* **292** 195–202 (1999); PSIPRED 4.0 (Predict Secondary Structure), http://bioinf.cs.ucl.ac.uk/psipred/, accessed on January 1, 2021.
[20] C. Mirabello and G. Pollastri, Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility, *Bioinformatics* **29** (16), 2056–2058 (2013); available at http://distillf.ucd.ie/porterpaleale/, accessed on January 1, 2021.
[21] L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass and M. J.E. Sternberg, the Phyre2 web portal for protein modeling, prediction and analysis, *Nature Protocols* **10** 845–858 (2015); the software is available at available at www.sbg.bio.ic.ac.uk/ phyre2/, accessed on January 1, 2021.
[22] S. Wang, S. Sun, Z. Li, R. Zhang and J. Xu, Accurate de novo prediction of protein contact map by ultra-deep learning model, *PLOS Comp. Biol.* (2017), DOI:10.1371/journal.pcbi.1005324. The software is available at available at http://raptorx.uchicago.edu/ContactMap/
[23] Genbank, available at https://www.ncbi.nlm.nih.gov/genbank/
[24] http://biomodel.uah.es/en/lab/cybertory/analysis/massager.htm
[25] https://web.expasy.org/translate/
[26] S. Dutta, I. V. Akey, C. Dingwall, K. H. Hartman, T. Laue, R. T. Nolte, J. F. Head and C. W. Akey, The crystal structure of nucleoplasmin-core: implications for histone binding and nucleosome assembly, *Molecular Cell* **8** 841–853 (2001).
[27] C. Sauter, J. Basquin and D. Suck, Sm-Like proteins in eubacteria: the crystal structure of the Hfq protein from Escherichia Coli, *Nucleic Acids* **31** 4091 (2003), https://www.rcsb.org/structure/1HK9.
[28] W. C.L. Lührmann, Spliceosome, structure and function, *Cold Spring Harbor Perspectives in Biology* **3** a003707 (2011).
[29] K. Irwin, M. Amaral and D. Chester, The Self-Simulation hypothesis interpretation of quantum mechanics, *Entropy* **22** 247 (2020).
[30] G. A. Jones, Maps on surfaces and Galois groups, *Math. Slov.* **47** 1–33 (1997).
[31] Planat, M. Geometry of contextuality from Grothendieck's coset space. *Quantum Inf. Process.* **2015**, *14*, 2563–2575.
[32] P. S. Aspinwall, K_3 surfaces and string duality, in C. Efthimiou and B. Greene, editors, "Fields, Strings and Duality, TASI 1996", pp 421–540, World Scientific, 1997 (Preprint hep-th/9611137).
[33] R. M. Koch and S. Ramgoolam, From matrix models and quantum fields to Hurwitz space and the absolute Galois group, Preprint 1002.1634 [hep-th].

† Université de Bourgogne/Franche-Comté, Institut FEMTO-ST CNRS UMR 6174, 15 B Avenue des Montboucons, F-25044 Besançon, France.
Email address: michel.planat@femto-st.fr

‡ Quantum Gravity Research, Los Angeles, CA 90290, USA
Email address: raymond@QuantumGravityResearch.org
Email address: Klee@quantumgravityresearch.org
Email address: Marcelo@quantumgravityresearch.org
Email address: Fang@QuantumGravityResearch.org