Graph Coverings for Investigating Non Local Structures in Proteins, Music and Poems

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

 doi:10.20944/preprints202108.0437.v1

GRAPH COVERINGS FOR INVESTIGATING NON LOCAL STRUCTURES IN PROTEINS, MUSIC AND POEMS

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

ABSTRACT. It is shown how the secondary structure of proteins, musical forms and verses of poems are approximately ruled by universal laws relying on graph coverings. In this direction, one explores the group structure of a variant of the SARS-Cov-2 spike protein and the group structure of apolipoprotein-H, passing from the primary code with amino acids to the secondary structures organizing the foldings. Then one look at the musical forms employed in the classical and contemporary periods. Finally, one investigates in much detail the group structure of a small poem in prose by Charles Baudelaire and that of the Bateau Ivre by Arthur Rimbaud.

1. INTRODUCTION

Let $\mathrm{rel}(x_1, x_2, \ldots, x_r)$ be the relation defining the finitely presented group $fp = \langle x_1, x_2, \ldots, x_r | \mathrm{rel}(x_1, x_2, \ldots, x_r) \rangle$ on r letters (or generators). We are interested in the conjugacy classes (cc) of subgroups of fp with respect to the nature of the relation rel. In a nutshell, one observes that the cardinality structure $\eta_d(fp)$ of conjugacy classes of subgroups of index d of fp is all the closer to that a the free group F_{r-1} on $r-1$ generators as the choice of rel contains more non local structure. To arrive at this statement, one experiments on protein foldings, musical forms and poems. The former case was first explored in [1].

The conjugacy classes subgroups of index d in the fundamental group of a base graph X are in one-to-one correspondence with the connected d-fold coverings of X, as it has been known for a long time [2, 3]. The derivation of such a result starts from an enumeration of integer partitions of d that satisfy

$$l_1 + 2l_2 + \ldots + dl_d = d,$$

a famous problem in analytic number theory [4, 5]. The number of such partitions is $p(d) = [1, 2, 3, 5, 7, 11, 15, 22 \cdots]$ when $d = [1, 2, 3, 4, 5, 6, 7, 8 \cdots]$.

The number of d-fold coverings of a graph X of first Betti number r is [3, p. 41]

$$\mathrm{Iso}(X; d) = \sum_{l_1+2l_2+\ldots+dl_d=d} (l_1! 2^{l_2} l_2! \ldots + d^{l_d} l_d!)^{r-1}.$$

Another interpretation of $\mathrm{Iso}(X; d)$ was found in [6, Eq. 12]. Taking a set of mixed quantum states comprising $r+1$ subsystems, $\mathrm{Iso}(X; d)$ corresponds to the stable dimension of degree d local unitary invariants. For two

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 August 2021 doi:10.20944/preprints202108.0437.v1

subsystems, $r = 1$ and such a stable dimension is $\mathrm{Iso}(X;d) = p(d)$. A table for $\mathrm{Iso}(X,d)$ with small d's is in [3, Table 3.1, p. 82] or [6, Table 1] .

Then, one needs a theorem derived by Hall in 1949 [7] about the number $N_{d,r}$ of subgroups of index d in F_r

$$N_{d,r} = d(d!)^{r-1} - \sum_{i=1}^{d-1} [(d-i)!]^{r-1} N_{i,r}$$

to establish that the number $\mathrm{Isoc}(X;d)$ of connected d-fold coverings of a graph X (alias the number of conjugacy classes of subgroups in the fundamental group of X) is as follows [3, Theorem 3.2, p. 84]

$$\mathrm{Isoc}(X;d) = \frac{1}{d} \sum_{m|d} N_{m,r} \sum_{l|\frac{d}{m}} \mu\left(\frac{d}{ml}\right) l^{(r-1)m+1},$$

were μ denotes the number-theoretic Möbius function.

Table 1 provides the values of $\mathrm{Isoc}(X;d)$ for small values of r and d [3, Table 3.2].

TABLE 1. The number $\mathrm{Isoc}(X;d)$ for small values of first Betti number r (alias the number of generators of the free group F_r) and index d. Thus the columns correspond to the number of conjugacy classes of subgroups of index d in the free group of rank r.

r	d=1	d=2	d= 3	d=4	d=5	d=6	d=7
1	1	1	1	1	1	1	1
2	1	3	7	26	97	624	4163
3	1	7	41	604	13753	504243	24824785
4	1	15	235	14120	1712845	371515454	127635996839
5	1	31	1361	334576	207009649	268530771271	644969015852641

We investigate three applications of the graph covering approach just introduced. In Section 2, it is shown that the secondary structures of two proteins (the spike protein of the SARS-Cov-2 virus and a glycoprotein playing a role in the immune system) approximately follow the theory just described, see [1] for our earlier work. In Section 3, the secondary structures are taken to be the musical forms. Then, in Section 4, the secondary structures in the verses of a poem are obtained from an encoding of the types of words (names, verbs, prepositions, etc).

In the latter two cases, like for proteins, our group theory applies reasonably well. The finitely presented groups $G = f_p$ that are of investigation in the present work may be characterized in terms of a first Betti number r. For a group G, r is the rank (the number of generators) of the abelian quotient $G/[G,G]$. To some extent, a group f_p whose first Betti number is r may be said to be close to the free group F_r since both of them have the same minimum number of generators.

 doi:10.20944/preprints202108.0437.v1

2. Graph coverings for proteins

2.1. **The D614G variant (minus RBD) of the SARS-CoV-2 spike protein.** As a first example of the application of our approach, let us consider the D614G variant (minus RBD: the receptor binding domain) of the SARS-CoV-2 spike protein. In the Protein Data Bank in Europe the name of the sequence is 6XS6 [8]. D614G is a missense mutation (a nonsynonymous substitution where a single nucleotide results in a codon that codes for a different amino acid). The mutation occurs at position 614 where G (glycine) has replaced D (aspartic acid) in many countries. It is found that G increases the transmission rate and correlates with the prevalence of loss of smell as a symptom of COVID-19, possibly related to a higher binding of the RBD to the ACE2 receptor: an enzyme attached to the membrane of heart cells. A picture of the secondary structures can be found in Figure 1.

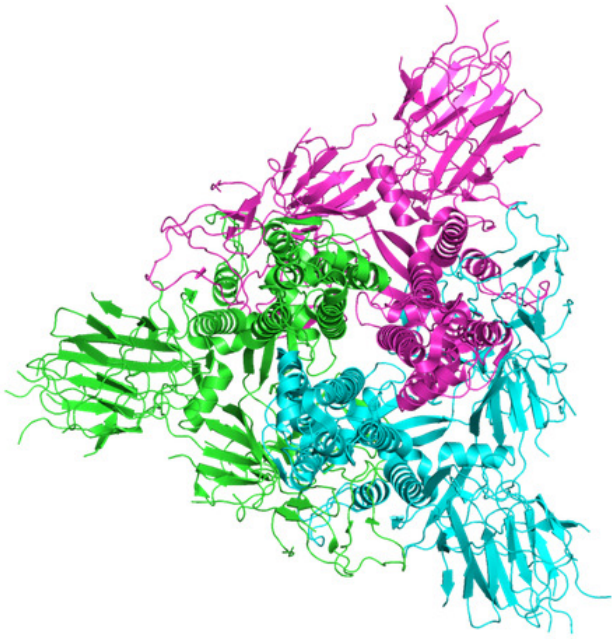

Figure 1. A picture of the secondary structure of D614G variant (minus RBD) of the SARS-CoV-2 spike protein found in the protein data bank in Europe [8] .

The D614G variant (minus RBD) of SARS-CoV-2 spike protein contains 786 aminoacids (aa) forming a (long) word as follows:

AYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHDNPVLPF...
AYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVV.
NTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKV...
FVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTV

Such a protein sequence, comprising 20 aminoacids as letters of the primary code, can be encoded in terms of secondary structures. Most of the time, for proteins, one makes use of three types of encoding that are segments of α helices (encoded with the symbol H), segments of β pleated

sheets (encoded by the symbol E) and segment of random coils (encoded by the segment C) [1, 9, 10].

But a finer structure may be obtained by making use of methods such as SST Bayesian method. A summary of the approach can be found in Reference [10].

We used a software prepared in [11] to get the following secondary structure

rel(H,E,C,G,I,T,4)=
CCCCCCCCCEEEEEECCCCCCCCEEEEECCCCCCCCCCCEEEEEECCCCCCCC...
HHHHHHHHCC444444CHHHHHHHHHHHHHHHHHHHHHCCCGGGGGHHHHH
HHIIIIICCCCCCCCCCCCCCCCCCTTTTTCCCCCCCCCCHHHHHHHHHHH...
CCCTTTTTCCCCCTTTTTCCCC44444EEEEEECC

where G means a 3_{10} helix, 4 means α-like turns, I means a right-handed π helix and T corresponds to unspecified turns.

For the group analysis, we slightly simplifies the problem by taking $4 = H$ that is by taking just one form of α turn so that the sequence is encoded with 6 letters only. Then, one further simplifies by taking $T = C$ to obtain a 5-letter encoding. One further simplifies by taking $I = H$, then by taking $G = H$ to get 4-letter and 3-letter encodings, respectively. The results are in Table 2.

TABLE 2. Group analysis of the D614G variant (minus RBD) of the SARS-CoV-2 spike protein. The bold numbers means that the cardinality structure of cc of subgroups of G fits the one of the free group F_{r-1} when the encoding makes use of r letters. In the last column, r is the first Betti number of the generating group f_p.

PDB 6XS6: AYTNSFTRGVYYPDKVFRSSVLHSTQDL ...	cardinality structure of cc of subgroups	r
6 letters H, E, C, G, I, T	[**1,31,1361,334576**]	5
5 letters H, E, C, G, I	[**1,15,235,14120**]	4
4 letters H, E, C, G	[**1,7,41 604**,14720]	3
3 letters H, E, C	[**1,3,7**,30,127, 926]	2

One observes that the cardinality structure of cc of subgroups of the finitely presented groups $fp = \langle H, E, C, G, I, T|rel\rangle, ..., fp = \langle H, E, C|rel\rangle$ fits the one of the free group F_{r-1} when the encoding makes use of $r = 6, 5, 4, 3$ letters. This in line with our results found in [1] on several kinds of proteins.

2.2. **The β-2-glycoprotein 1 or apolipoprotein-H.** Our second example deals about a protein playing an important role in the immune system [12]. In the protein data bank, the name of the sequence is 6V06 [13]. It contains 326 aa. All models predict secondary structures mainly comprising β-pleated sheets and random coils and sometimes short segments of α-helices.

One observes in Table 3 that the cardinality structure of cc of subgroups of the finitely presented groups $fp = \langle H, E, C|rel\rangle$ fits reasonably well the one of the free group F_2 on two letters for the first three models but not for the RAPTORX model. In one case (with the PORTER model [14]), all first

TABLE 3. Group analysis of apolipoprotein-H (PDB 6V06). The bold numbers means that the cardinality structure of cc of subgroups of f_p fits the one of the free group F_3 when the encoding makes use of 2 letters. The first model is the one used in the previous subsection [11] where we took $4 = H$ and $T = C$. The other models of secondary structures with segments E, H and C are from softwares PORTER, PHYRE2 and RAPTORX. The references to these softwares may be found in our recent paper [1]. The notation r in column 3 means the first Betti number of f_p.

PDB 6V06: GRTCPKPDDLPFSTVVPLKTFYEPG...	cardinality structure of cc of subgroups	r
Konagurthu	[**1,3,7,26**,218,2241]	2
PORTER	[**1,3,7,26,97,624**]	.
PHYRE2	[**1,3,7,26**,157,1046]	.
RAPTORX	[**1**,7,17,134,923, 13317]	3

FIGURE 2. A picture of the secondary structure of the apolipoprotein-H obtained with the software [11].

six digits fit those of F_2 and higher order digits could not be reached. The reader may refer to our paper [1] where such a good fit could be obtained for the sequences in the arms of the protein complex Hfq (with 74 aa). This complex with 6-fold symmetry is known to play a role in DNA replication.

A picture of the secondary structure of the apolipoprotein-H obtained with the software of Ref. [11] is displayed in Table 2.

3. Graph coverings for musical forms

Nobody would not accept that the structure determines the beauty in art. We provide two examples of this relationship, first by studying musical forms, then by looking at the structure of verses in poems. Our approach encompasses the orthodox view of periodicity or quasi-periodicity inherent to such structures. Instead of that, the non local character of the structure is investigated thanks to a group with generators given by the allowed generators $x_1, x_2, \cdots, x_r$ and a relation rel determining the position of such successive generators, as we did for the secondary structures of proteins.

3.1. The sequence Isoc$(X;1)$, the Golden ratio and more.

The Fibonacci sequence. As shown in Table 1, the sequence Isoc$(X;1)$ only contains 1 in its entries and it is tempting to associate this sequence to the most irrational number, the Golden ratio $\phi = (\sqrt{5}-1)/2$ through the continued fraction expansion $\phi = 1/(1+1/(1+1/(1+1/(1+\cdots)))) = [0;1,1,1,1,\cdots)$.

Let us now take a two-letter alphabet (with letters L and S) and the Fibonacci words w_n defined as $w_1 = S$, $w_2 = L$, $w_n = w_{n-1}w_{n-2}$. The sequence of Fibonacci words w_n is as follows

$$S, L, LS, LSL, LSLLS, LSLLSLSL, LSLLSLSLLSLLS, LSLLSLSLLSLLSLSLLSLSL, \cdots$$

and its length corresponds to the Fibonacci numbers $1, 1, 2, 3, 5, 8, 13, 21, \cdots$.

Then, one can check that the finitely-presented group $f_p(n) = \langle S, L | w_n \rangle$ whose relation is a Fibonacci word w_n possesses a cardinality sequence of subgroups $[1, 1, 1, 1, 1, 1, 1, 1 \cdots)$ equal to Isoc$(X;1)$, up to all computable orders, despite the fact that the groups $f_p(n)$ are not the same.

It is straightforward to check that the first Betti number r of $f_p(n)$ is 1 as expected.

The period doubling cascade. Other rules lead to a Betti number $r = 1$ and the corresponding sequence Isoc(X;1). Let consider the period-doubling cascade in the logistic map $x_{l+1} = 1 - \lambda x_l^2$. Period doubling can be generated bu repeated use of the substitutions $R \to RL$ and $L \to RR.$, so that the sequence of period doubling is [15]

$$R, L, RL, RLR^2, RLR^3LRL, RLR^3LRLRLR^3LR^3, RLR^3LRLRLR^3LR^3LR^3LRLRLR^3LRLRL, \cdots$$

and the corresponding finitely presented groups also have first Betti number equal to 1.

Musical forms of the classical age. Going to musical forms, the ternary structure $L - S - L$ (most commonly denoted $A - B - A$) corresponding to the Fibonacci word w_4 is a Western instrumental genre notably used in sonatas, symphonies and string quartets. The basic elements of sonata form are the exposition A, the development B and recapitulation A. While the musical form $A - B - A$ is symmetric, the Fibonacci word $A - B - A - A - B$ corresponding to w_5 is asymmetric and used in some songs or ballades from the Renaissance.

In a closely related direction, it was shown that the lengths a and b of sections A and B in all Mozart's sonata movements are such that the ratio $b/(a+b) \approx \phi$ [16].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 August 2021 doi:10.20944/preprints202108.0437.v1

3.2. **The sequence Isoc$(X;2)$ in twentieth century music and jazz.** In the 20th century, musical forms escaping the classical channels were created. With the Hungarian composer Béla Bartók, a musical structure known as arch form is created. The arch form is a sectional structure for a piece of music based on repetition, in reverse order, of all or most musical sections such that the overall form is symmetric, most often around a central movement. Formally, it looks like $A-B-C-B-A$. A well known composition of Bartok with this structure is *Music for strings, percussion and celesta* [17]. In Table 4, it is shown that the cardinality sequence of cc of subgroups of the group generated with the relation rel=$ABCBA$ corresponds to Isoc$(X;2)$ up to the higher index 9 that we could check with our computer. A similar result is obtained with the symmetrical word $ABACABA$.

TABLE 4. Group analysis of a few musical forms whose structure of subgroups, apart from exceptions, is close to Isoc$(X;d)$ with $d=2$ (at the upper part of the table) or $d=3$ (at the lower part of the table). Of course, the forms A-B-C and A-B-C-D have the cardinality sequence of cc of subgroups exactly equal to Isoc$(X;2)$ and Isoc$(X;3)$, respectively.

musical form	ref	cardinality structure of cc of subgroups	r
A-B-C-B-A	arch, Belá Bartók	[**1,3,7,26,97,624,4163,34470,314493**]	2
A-B-A-C-A-B-A	.	.	.
A-B-A-C-A, A-B-A-C-A-B-A	rondo	.	.
A-B-A-C		.	.
A-A-B-C-C	Haydn [19], djanba [20, Fig. 9.8]	.	.
A-A-A-A-B-B-A-A-C-C-A-A	twelve-bar blues, standard	[1,7,14,109,396,3347,19758,287340]	3
A-A-A-A-B-B-A-A-C-B-A-A	twelve-bar blues, variation 1	[**1,3,7,26,97,624,4163,34470,314493**	2
A-A-A-A-B-B-A-A-B-C-A-C	twelve-bar blues, variation 2	[**1,3,7,26**,127, 799, 5168, 42879]	.
A-B-C	Isoc$(X;2)$	[**1,3,7,26,97,624,4163,34470,314493**]	2
A-A-B-B-C-C-D-D	pot pourri	[**1,15**,82,1583,30242]	4
A-B-A-C-A-D-A	rondo	[**1,7,41,604,13753,504243**]	3
A-B-C-D	Isoc$(X;3)$	[**1,7,41,604,13753,504243,24824785**]	3

Our second example is a musical form known as twelve-bar blues [18], one of the most prominent chord progressions in popular music and jazz. In this context, the notation A is for the tonic, B is for the subdominant and C is for the dominant, each letter representing one chord. In twelve-bar blues, there are twelve chords arranged as in the first column of Table 4. One observes that the standard twelve-bar blues is away in structure from the sequence of Isoc$(X;2)$. But variations 1 and 2 have a structure close to Isoc$(X;2)$. In the former case, the first 9 orders lead to the same digit in the sequence.

Our third example is the musical form A-A-B-C-C. Notably, it is found in the *Slow movement from Haydn's 'Emperor' quartet Opus 76, N3* [19] (Fig. 3), much sooner than the contemporary period. See also Ref. [20] for the frequent occurrence of the same musical form in djanba songs at Wadeye. As in the aforementioned examples, the cardinality sequence of cc of subgroups of the group built with rel=AABCC corresponds to Isoc$(X;2)$ up to the highest index 9 that we could reach in our calculations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 August 2021 doi:10.20944/preprints202108.0437.v1

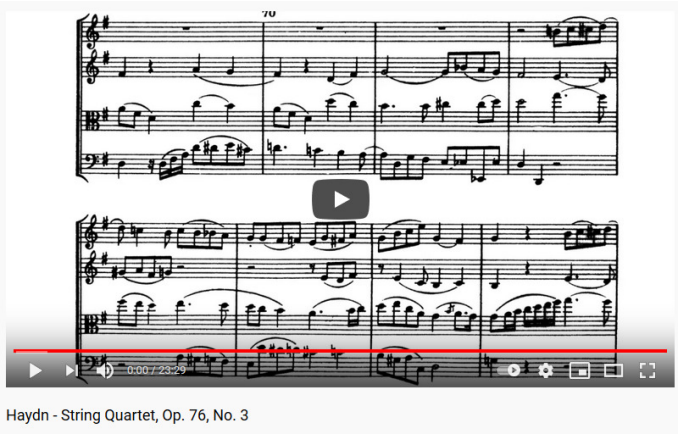

FIGURE 3. Slow movement from Haydns Emperor quartet Opus 76, N3.

Further musical forms with 4 letters A, B, C, and D and their relationship to Isoc$(X;3)$ are provided in the lower part of Table 4.

Not surprisingly, the rank r of the abelian quotient of $f_p = \langle A, B, C|\text{rel}(A, B, C)\rangle$ is found to be 2 when the cardinality structure fits that Isoc$(X;2)$ in Table 4. Otherwise the rank is 3. Similarly, the rank r of the abelian quotient of $f_p = \langle A, B, C, D|\text{rel}(A, B, C, D)\rangle$ is found to be 3 when the cardinality structure fits that Isoc$(X;3)$ in Table 4. Otherwise the rank is 4.

4. Graph coverings for prose and poems

4.1. **Graph coverings for prose.** Let us perform a group analysis of a long sentence in prose. We selected a text by Charles Baudelaire [21]:

Le gamin du cleste Empire hsita dabord; puis, se ravisant, il rpondit: "Je vais vous le dire ". Peu dinstants aprs, il reparut, tenant dans ses bras un fort gros chat, et le regardant, comme on dit, dans le blanc des yeux, il affirma sans hsiter: "Il nest pas encore tout fait midi." Ce qui tait vrai.

In Table 5, the group analysis is performed with 3, 4 or 5 letters (in the upper part) and compared to random sequences with the same number of letters (in the lower part).

The text of the sentence is first encoded with three letters (H for names and ajectives, E for verbs and C otherwise), in that case one observes that the subgroup structure has cardinality close to that of a free group F_2 on two letters up to index 3. If one adds one letter A for the prepositions in the sentence (in addition to H, E and C), then the subgroup structure has cardinality close to that of a free group F_3 on three letters. And if adverbs B are also selected, then the subgroup structure is close to that of the free group F_4. In all three cases, one can check that the similarity holds up to index 3 and that the cc of subgroups are the same than in the corresponding free groups. The first Betti numbers of the generating groups are 2, 3 and 4 as expected.

In Table 5, one also computed the cardinality structure of cc of subgroups of small index obtained from a random sequence of 250 letters (like the number of letters in the previously studied sentence of the small poem in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 August 2021 doi:10.20944/preprints202108.0437.v1

prose). One took 10 runs with random sequence having 3, 4 or 5 letters. One clearly sees that the cardinality structure of cc of subgroups for the cases with 4 or 5 letters tends to align to that of the free group F_{r-2} (not F_{r-1}). The 3-letter case is the most random one and does not correspond to F_1 (or F_2), in most runs.

Our conclusion is that the considered prose sequence contains structure close to that of F_r when one selects $r+1$ letters for the encoding of the sentence, a result that is similar to that we already found in the group analysis of proteins in Sect. 2 and musical forms in Sect. 3.

TABLE 5. Group analysis of an excerpt of small poem in prose *Le vieux saltimbanque* by Charles Baudelaire. The text is split into segments encoded by the symbol H (for names and adjectives), E (for verbs), A for prepositions, B for adverbs, or C (for the other types: conjonctions, punctuation marks and so on). The cardinality structure of cc of subgroups of a small index is compared to the one obtained with 10 runs of a sequence of words of similar length (i.e. the length 250) with the corresponding number of letters.

Le gamin du cleste Empire ... Ce qui tait vrai.	card. seq. of cc of subgroups	r
3 letters: rel=$C^2H^5C^2H^7H^6E^6C^7CC^4CC^2E^8C\cdots$	[**1,3,7**,34,131]	2
4 letters: rel=$C^2H^5A^2H^7H^6E^6C^7CC^4CC^2E^8C\cdots$	[**1,7,41**,636,14364]	3
5 letters: rel=$C^2H^5A^2H^7H^6E^6B^7CB^4CC^2E^8C\cdots$	[**1,15,235**,14376,.]	4
[Random[1,3]: i in [1..250]] (10 runs)	[**1**,1,1,2,4,4]	1
	[**1,3**,2,9,5,20]	2
	[**1,3**,1,6,6,15]	.
	[**1,3,7**,30,124,987]	.
	[**1**,7,17,126,323,2445]	3
	etc	
Isoc(X;2)	[**1,3,7,26,97,624**]	2
[Random[1,4]: i in [1..250]] (10 runs)	[**1,3,7**,30,.] (×3)	2
	[**1,3,10**,51,.] (×3)	.
	[**1,3,7,26**,457]	.
	[**1,3**,10,39,.]	.
	[**1,3**,13,52,.]	.
	[**1**,7,20,143,.]	3
Isoc(X;3)	[**1,7,41,604,13573**]	3
[Random[1,5]: i in [1..250]] (10 runs)	[**1,7,41**,620,.] (×3)	3
	[**1,7,41**,636,.] (×3)	.
	[**1,7,41,604**,.] (×2)	.
	[**1,7,41**,668,.]	.
	[**1,7**,50,819,.]	.
Isoc(X;4)	[**1,15,235,14120,1712845**]	4

4.2. **Graph coverings for poems.** In poems, the verses are generally of smaller length than that for a sentence in prose. One selected the first strophe of the poem *Le Bateau Ivre* by Arthur Rimbaud and an English translation. The poem may be found on a wall in Paris, see Figure 4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 August 2021 doi:10.20944/preprints202108.0437.v1

The verses in the strophe have about 35 letters. One compares the group structure of the four verses in the first strophe to that of random sequences of length 35 in Table 6 (when the encoding is with 3 letters H, E and C) and in Table 7 (when the encoding is with 4 letters H, E, C and A). Adverbs are too rare in verses of such a small length so that we did not considered the 5-letter case.

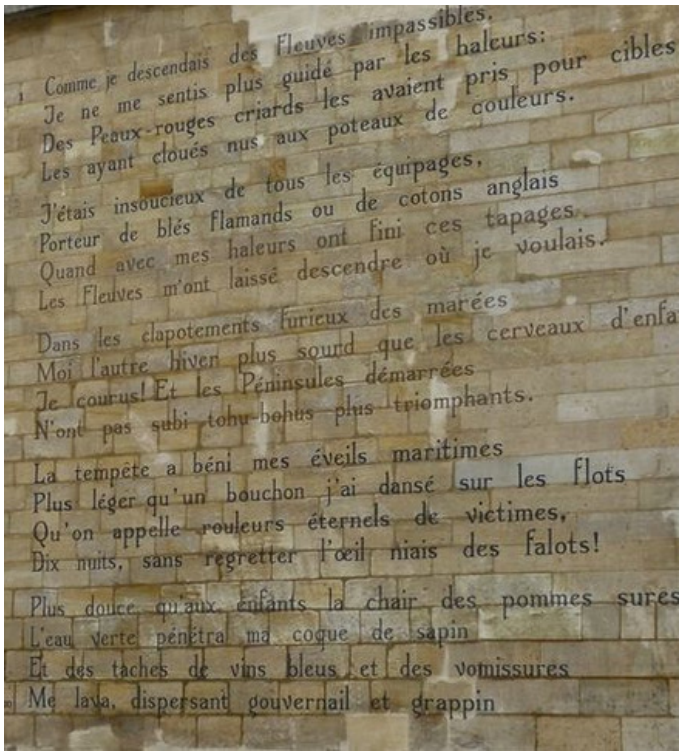

FIGURE 4. Part of the poem *Le Bateau Ivre* of Arthur Rimbaud on a wall (Rue Férou) in Paris.

Let us first look at the 3-letter case in Table 6. Apart from the first verse in the strophe, the structure of the poem is very close to that of F_2 up to the index 6 (for the second verse) and up to the index 7 (for verses 3 and 4). Higher order indices could not be reached in our calculations. For the English translation, the closeness to F_2 holds as well but is not so perfect. It is not so surprising since the poem was originally composed in French. For a French translation of a poem in English one would have obtained a similar (small) discrepancy to the group structure to F_2. We looked at the cardinality structure of cc of subgroups by taking random sequences of length 35 in 10 runs and we observe that the closeness to F_2 is much less than in the case of the poem.

Then, all that has been found for the group structure with 3 letters can also be obtained for the group structure with 4 letters in Table 7 but the closeness is to F_3 (not F_2), as expected.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 August 2021 doi:10.20944/preprints202108.0437.v1

TABLE 6. Group structure of the poem *Le Bateau Ivre'* (*The Drunken Boat*) by Arthur Rimbaud. Only the first strophe (that has four lines) is analyzed, firstly in its original form, then in an English translation. Each line is split into segments encoded by the symbol H (for names and adjectives), E (for verbs) or C (for the other types: conjonctions, adverbs, prepositions, punctuation marks and so on). The group relation is displayed for the first line only.) The cardinality structure of cc of subgroups of a small index is compared to the one obtained with 10 runs of a sequence of random 3-letter words of similar length (i.e. the length 35).

Comme je descendais des fleuves impassibles,	[**1,1,7,**17,114, 1395,36973]	1
rel=$C^4C^2E^{10}C^3H^7H^{11}C$		
Je ne me sentis plus guid par les haleurs:	[**1,3,7,26,97, 624**,4171]	2
Des Peaux-Rouges criards les avaient pris pour cibles	[**1,3,7,26,97, 624,4163**]	.
Les ayant clous nus aux poteaux de couleurs.	[**1,3,7,26,97, 624,4163**]	.
As I was floating down unconcerned rivers	[**1,3,7,26,97, 624,4163,34470**]	2
rel=$C^2 * C * E^3 * E^8 * C^4 * E^11 * H^6$		
I now longer felt myself steered by the haulers:	[**1,3,7,26**,101, 656,4227]	2
Gaudy Redskins had taken them for targets	[**1,3,7,26,97, 624,4163**,324935]	.
Nailing them naked to coloured states.	[**1,3,7**,42,202, 1682,9204]	.
[Random[1,3]: i in [1..35]]	[**1,3,7**,30, .] (×3)	2
(10 runs)	[**1,3,7,26**, .](×3)	.
	[**1,3,7**,.,.,]	.
	[**1,3,10,.,.**](×2)	.
	[**1,3**,13,.,.]	.
Isoc(X;2)	[**1,3,7,26,97, 624,4163,34470**]	2

5. CONCLUSION

The graph covering approach has been shown to be useful for understanding how complex structures are encoded in nature and in art. For proteins, there exists a primary encoding with 20 amino acids as letters and the secondary encoding determines the folding of proteins in the 3-dimensional space. This is useful for recognizing the relationship between the structure and function of the protein. We took examples based on a presently hot topic: a variant of the SARS-Cov-2 spike protein and the alipoprotein-H. For music, the secondary structures are called musical forms and the choice of them determines the type of music. For poems, we took the French (or English) alphabet with 26 letters but many other alphabets may be used for the application of our approach. The secondary structures are defined from the encoding of the words (names, verbs and so on).

It may be interesting to speculate about the possible existence of a primary code and a secondary code in other fields, e.g. in physics at the elementary level like in particle physics and quantum gravity [24]. According to the experience of the authors of this paper, the structure has much to do with complete quantum information. The reader may consult paper [22] about particle mixings or [1, 23] about the genetic code in which finite groups are the players. Here we are dealing with infinite groups so that the

TABLE 7. The same as in Table 6 but each line is split into segments encoded by the symbol H (for names and adjectives), E (for verbs), A for prepositions, or C (for the other types: conjonctions, adverbs, punctuation marks and so on). The cardinality structure of cc of subgroups of a small index is compared to the one obtained with 10 runs of a sequence of random 4-letter words of similar length (i.e. the length 35).

Comme je descendais des fleuves impassibles,	[**1,7,41,604,13753**]	3
rel=$C^4C^2E^{10}A^3H^7H^{11}C$		
Je ne me sentis plus guid par les haleurs:	[**1,7,41,604,13753**]	.
Des Peaux-Rouges criards les avaient pris pour cibles	[**1,7,41,604,13753**]	.
Les ayant clous nus aux poteaux de couleurs.	[**1,7,41,604,13753**]	.
As I was floating down unconcerned rivers	[**1,7**,59,1386,27011]	3
rel=$C^2CE^3E^8A^4E^{11}H^6$		
I no longer felt myself steered by the haulers:	[**1,7,41,604,13753**]	.
Gaudy Redskins had taken them for targets	[**1,7**,50,1763,51582]	.
Nailing them naked to coloured states.	[**1,7**,59,1002,18671]	.
[Random[1,4]: i in [1..35]]	[**1,7**,50,755,.] (×2)	3
(10 runs)	[**1,7,41,604**,.] (×3)	.
	[**1,7,41**,.,.](×2)	.
	[**1,7**,50,739,.](×2)	.
	[**1,7**,59,.,.]	.
Isoc(X;3)	[**1,7,41,604,13753**]	3

representation theory of finite groups (with characters) has to be defined on finitely-presented groups (most of the time of infinite cardinality). This will be done in a next work.

AUTHOR CONTRIBUTIONS

Conceptualization, M.P., F.F. and K.I.; methodology, M.P. and R.A.; software, M.P.; validation, R.A., F.F. and M. M.A..; formal analysis, M.P. and M. M.A.; investigation, M.P., F.F. and M. M.A.; writing–original draft preparation, M.P.; writing–review and editing, M.P.; visualization, F.F. and R.A.; supervision, M.P. and K.I.; project administration, K.I..; funding acquisition, K.I. All authors have read and agreed to the published version of the manuscript.

FUNDING

Funding was obtained from Quantum Gravity Research in Los Angeles, CA.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Planat, M.; Aschheim, R.; Amaral, M. M; Fang, F.; Irwin, K. Quantum information in the protein codes, 3-manifolds and the Kummer surface, *Symmetry* **2020**, 13, 1146.
[2] Mednykh, A. Counting conjugacy classes of subgroups in a finitely generated group, *J. Algebra* **2008**, 320, 2209–2217.

[3] Kwak, J. H.; Nedela R., Graphs and their coverings, *Lecture Notes Series* **2007**, 17, 118 pp.
[4] Hardy G. H.; Wright E. M., *An introduction to the theory of numbers*, 6th ed.; Oxford University Press: Oxford, UK, 2008; pp. 361–392.
[5] Planat, M. Quantum $1/f$ noise in equilibrium: from Planck to Ramanujan, *Physica A* **2003**, 318, 371.
[6] Vrna, P. On the algebra of local unitary invariants of pure and mixed quantum states, *J. Phys. A: Math. Theor.* **2011**, 44, 225304.
[7] Hall Jr M.. Subgroups of finite index in free groups, *Can. J. Math.* **1949**, 1, 187–190.
[8] SARS-CoV-2 Spike D614G variant, minus RBD, in Protein Data Bank in Europe, Bringing Structure to Biology, available at https://www.ebi.ac.uk/pdbe/entry/pdb/6xs6, accessed on 1 May 2021.
[9] Dang, Y.; Gao, J.; Wang, J.; Heffernan, R.; Hanson, J.; Paliwal, K.; Zhou, Y. Sixty-five years of the long march in protein secondary structure prediction: The final strech? *Brief. Bioinform.* **2018**, *19*, 482–494.
[10] Protein secondary structure,
available at https://en.wikipedia.org/wiki/Protein_secondary_structure, accessed on 1 May 2021.
[11] Konagurthu,A. S.; Lesk A. M.; Allison L. Minimum message length inference of secondary structure fromprotein coordinate data. *Bioinformatics* **2012**, 28, i97-i105. The software is at https://lcb.infotech.monash.edu/sstweb2/Submission_page.html, accessed on 1 May 2021.
[12] McDonnell T.; Wincup C.; Bucholz I.; Pericleous C.; Giles I.; Ripoll V.; Cohen H.; Delcea M.; Rahman A. The role of beta-2-glycoprotein I in health and disease associating structure with function: More than just APS. *Blood Rev.* **2020**, *39*, 100610.
[13] Crystal structure of Beta-2 glycoprotein I purified from plasma (pB2GPI), available at https://www.rcsb.org/structure/6V06 , accessed on 1 May 2021.
[14] Mirabello, C.; Pollastri, G. Porter, PaleAle 4.0: High-accuracy prediction of protein secondary structure and relative solvent accessibility. *Bioinformatics* **2013**, *29*, 2056–2058.
[15] Flicker, F. Time quasilattices in dissipative dynamical systems. *SciPost Phys.* **2018**, *5*, 001.
[16] Putz, J. F. The Golden section and the piano sonatas of Mozart. *Mathematics Magazine* **1995**, *68*, 275–282.
[17] Music for strings, percussion and celesta,
available at https://en.wikipedia.org/wiki/Music_for_Strings,_Percussion_and_Celesta, accessed on 1 May 2021.
[18] Twelve-bar blues, available at https://en.wikipedia.org/wiki/Twelve-bar_blues, accessed on 1 May 2021.
[19] Haydn - String Quartet, Op. 76, No. 3,
available at https://www.youtube.com/watch?v=qoWdtGUe5fc, accessed on 1 May 2021.
[20] Barwick L. Musical form and style in Murriny Patha djanba songs at Wadeye (Northern Territory, Australia, available at https://core.ac.uk/download/pdf/41240492.pdf, accessed on 1 May 2021.
[21] Baudelaire, C. Le vieux saltimbanque, in *Petits poèmes en prose* (1869).
[22] Planat, M.; Aschheim, R.; Amaral, M.M.; Irwin, K. Informationally complete characters for quark and lepton mixings. *Symmetry* **2020**, *12*, 1000.
[23] Planat, M.; Chester, D.; Aschheim, R.; Amaral, M.M.; Fang, F.; Irwin, K. Finite groups for the Kummer surface: The genetic code and quantum gravity. *Quantum Rep.* **2021**, *3*, 68–79.
[24] Irwin, K.; Amaral, M.; Chester, D. The Self-Simulation hypothesis interpretation of quantum mechanics. *Entropy* **2020**, *22*, 247.

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

 doi:10.20944/preprints202108.0437.v1

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

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