



Article

# Group Theory of Syntactical Freedom in DNA Transcription and Genome Decoding

Michel Planat <sup>1,\*</sup>, Marcelo M. Amaral <sup>2,†</sup>, Fang Fang <sup>2,†</sup>, David Chester <sup>2,†</sup>, Raymond Aschheim <sup>2,†</sup> and Klee Irwin <sup>2,†</sup>

<sup>1</sup> Institut FEMTO-ST CNRS UMR 6174, Université de Bourgogne-Franche-Comté, F-25044 Besançon, France  
<sup>2</sup> Quantum Gravity Research, Los Angeles, CA 90290, USA; marcelo@QuantumGravityResearch.org (M.M.A.); fang@QuantumGravityResearch.org (F.F.); davidc@QuantumGravityResearch.org (D.C.); raymond@QuantumGravityResearch.org (R.A.); klee@QuantumGravityResearch.org (K.I.)  
\* Correspondence: michel.planat@femto-st.fr  
† These authors contributed equally to this work.

**Abstract:** Transcription factors (TFs) are proteins that recognize specific DNA fragments in order to decode the genome and ensure its optimal functioning. TFs work at the local and global scales by specifying cell type, cell growth and death, cell migration, organization and timely tasks. We investigate the structure of DNA-binding motifs with the theory of finitely generated groups. The DNA ‘word’ in the binding domain—the motif—may be seen as the generator of a finitely generated group  $F_{dna}$  on four letters, the bases A, T, G and C. It is shown that, most of the time, the DNA-binding motifs have subgroup structures close to free groups of rank three or less, a property that we call ‘syntactical freedom’. Such a property is associated with the aperiodicity of the motif when it is seen as a substitution sequence. Examples are provided for the major families of TFs, such as leucine zipper factors, zinc finger factors, homeo-domain factors, etc. We also discuss the exceptions to the existence of such DNA syntactical rules and their functional roles. This includes the TATA box in the promoter region of some genes, the single-nucleotide markers (SNP) and the motifs of some genes of ubiquitous roles in transcription and regulation.

**Keywords:** DNA transcription factors; finitely generated groups; aperiodic order



**Citation:** Planat, M.; Amaral, M.M.; Fang, F.; Chester, D.; Aschheim, R.; Irwin, K. Group Theory of Syntactical Freedom in DNA Transcription and Genome Decoding. *Curr. Issues Mol. Biol.* **2022**, *44*, 1417–1433. <https://doi.org/10.3390/cimb44040095>

Academic Editors: Chiara Rosso and Antonio Francavilla

Received: 21 February 2022

Accepted: 20 March 2022

Published: 22 March 2022

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In his recent paper, Klee Irwin writes “Reality would be non-deterministic, not because it is random, but because it is a code—a finite set of irreducible symbols and syntactical rules. Our definition of information is meaning conveyed by symbolism. And expressions of code or language are strings of symbols allowed by syntax—ordering rules with syntactical freedom.” [1].

Our last papers focused on the relevance of free groups in the encoding of the secondary structure of proteins [2] and in the encoding of tonal music and poetry [3].

In this present contribution, the concept of syntactical freedom is associated with that of a free group in the encoding of strings of symbols—the motifs of DNA transcription. It is shown for the first time that most transcription factors, but not all, have motifs where the DNA letters in the motif form a finitely generated group whose structure is close to a free group. The exceptions rely either on a specific functional role of the DNA sequence under investigation or a potential dysfunction in the transcription of the gene, resulting in disease.

A few definitions in the domain of genetics that we use in the paper are as follows:

**Sequence motif** An amino-acid sequence pattern that is related to a biological function or a gene. The motif is sometimes called a ‘consensus sequence’.

**DNA-binding domain** A folded protein domain that contains a structural motif that recognizes double- or single-stranded DNA.

**Transcription factor** A sequence-specific DNA-binding factor, or transcription factor, is a protein that controls the rate of transcription of a gene from DNA to messenger RNA by binding to a specific DNA sequence. There are approximately 1600 binding domains in the human genome that function as transcription factors. Classes of DNA-binding domains of transcription factors also exist. The most common are zinc-coordinating DNA-binding domains, helix-loop-helix or helix-turn-helix motifs, basic leucine zipper domains and homeobox domains (playing critical roles in the regulation of development). A classification of human transcription factors and their structural motifs is in References [4–6].

**Exon** A part of a gene that encodes a part of the mature RNA produced by that gene after removing all introns (the non-coding regions of RNA transcript) by RNA splicing.

**Promoter** A sequence of DNA in which proteins initiate the transcription of a single RNA from the DNA downstream of it. The TATA box is a sequence found in the core promoter region of some genes in archaea and eukaryotes.

**Zinc finger** A small protein structural motif containing one or more zinc ions in order to stabilize the protein fold.

**Protein isoform** A set of highly similar proteins may originate from a single gene. This process is regulated by the alternative splicing of mRNA. In this process, particular exons of a gene may be included within or excluded from the final, processed messenger RNA (mRNA) produced from that gene. Alternative splicing and the multi-exonic genes are a common feature in eukaryotes.

## 2. Materials and Methods

### 2.1. Finitely Generated Groups

In the domain of algebra, a group is a set equipped with an operation between every pair of elements of the set that is associative, has an identity element, and every element has an inverse. The elements of a group may be numbers or other structures.

One familiar group is the set  $\mathbb{Z}$  of relative integers, together with the addition as the group law. Notice that  $\mathbb{Z}$  has infinite elements. The addition in  $\mathbb{Z}$  is commutative, not a general property for other groups.

The groups under consideration in this paper have elements that are words on four letters A, T, G and C—the bases in DNA—and the group law is the product. In the context of transcription factors (TFs), a motif fully describes the group. For example, taking only two letters/bases A and T and a relation/motif  $\text{rel}(A,T) = \text{TATA}$ , the group is of the finitely generated type and the mathematical notation is  $f_p = \langle A, T | \text{TATA} = I \rangle$ , in which the motif equals the identity element  $I$ . Infinite elements exist in  $f_p$ . The group  $f_p$  is non-commutative and contains a wealth of symmetries. One useful method for investigating them consists of two important concepts.

The first one is the concept of a free group on two generators  $F_2 = \langle A, T \rangle$  with no explicit relation (but not forgetting the relations following from the axioms of a group:  $AA^{-1} = TT^{-1} = I$ ). The second concept is that of a conjugacy class of subgroups. Two elements  $a$  and  $b$  belong to the same class if they are conjugate, meaning that  $a$  and  $b = g^{-1}ag$  belong to the same class for every  $g \in f_p$ .

The finitely generated group with the relation  $\text{rel}(A,T) = \text{TATA}$  may be characterized by the number of conjugacy classes at each index  $d \in [1, 2, 3, \dots]$  with the sequence  $\eta_d = [1, 3, 3, 10, 15, 56, 131, 462, \dots]$ . A signature of the group is also shared by the group of ‘dessins d’enfants’ in [7]—which the reader may consult for an introduction to the field of finitely generated groups and the related graphs, topology and geometry.

### 2.2. Free Groups and Their Conjugacy Classes

Let  $F_r$  be the free group on  $r$  generators. Following a theorem derived by Hall in 1949 [8], the number  $N_{d,r}$  of subgroups of index  $d$  in  $F_r$  is

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

leading to the number  $\text{Isoc}(X; d)$  of connected  $d$ -fold coverings of a graph  $X$  (i.e., the number of conjugacy classes of subgroups in the fundamental group of  $X$ ) is as follows [9] (Theorem 3.2, p. 84),

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

where  $\mu$  denotes the number-theoretic Möbius function.

Table 1 provides the values of  $\text{Isoc}(X; d)$  for small values of  $r$  and  $d$ , see [2] and ([9] [Table 3.2]).

**Table 1.** The number of conjugacy classes of subgroups of index  $d$  in the free group of rank  $r$  [9].

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	13,753	504,243	24,824,785
4	1	15	235	14,120	1,712,845	371,515,454	127,635,996,839
5	1	31	1361	334,576	207,009,649	268,530,771,271	644,969,015,852,641

We are interested in the cardinality sequence (card seq) of conjugacy classes for subgroups of a finitely generated group  $f_p$  with a relation (rel) given by the sequence motif. Most of the time, the DNA motif in the transcription factor is close to that of a free group  $F_r$ , with  $r + 1$  being the number of distinct bases involved in the motif. However, the finitely generated group  $f_p = \langle x_1, x_2 | \text{rel}(x_1, x_2) \rangle$ , or  $f_p = \langle x_1, x_2, x_3 | \text{rel}(x_1, x_2, x_3) \rangle$ , or  $f_p = \langle x_1, x_2, x_3, x_4 | \text{rel}(x_1, x_2, x_3, x_4) \rangle$  (where the  $x_i$  are taken in the four bases A, T, G, C and rel is the motif) is not the free group  $F_1 = \langle x_1, x_2 | x_1 x_2 \rangle$ , or  $F_2 = \langle x_1, x_2, x_3 | x_1 x_2 x_3 \rangle$ , or  $F_3 = \langle x_1, x_2, x_3, x_4 | x_1 x_2 x_3 x_4 \rangle$ . The closeness of  $f_p$  to  $F_r$  can be checked in the finite range of indices of the card seq.

### 2.3. Content of the Paper

The structure of the TATA box in the core promoter region of many eukaryotes is not close to that of a free group. Remarkably, the card seq of  $f_p$  for the TATA box is close to that of the Hecke group  $H_q = \langle x_1, x_2 | x_1^2 = x_2^q \rangle$  [10]. The case  $q = 3$  corresponds to the modular group  $PSL(2, \mathbb{Z})$ , which is the fundamental group of the trefoil knot manifold  $K3a1 = 3_1$ . The Hecke group  $H_q$ , with  $q \geq 3$ , is the discrete group generated by  $z \rightarrow -1/z$ ,  $z \rightarrow z + \lambda_q$ , where  $\lambda_q = 2 \cos(\pi/q)$  with  $\lambda_3 = 1$ ,  $\lambda_4 = \sqrt{2}$ ,  $\lambda_5 = (1 + \sqrt{5})/2$ ,  $\lambda_6 = \sqrt{3}$ , etc. In Section 3.1, it is shown that the card seq for motifs in the standard TATA box corresponds to  $H_3$  or  $H_4$  and that, in the case of a Gilbert’s syndrome, it is only approximate or corresponds to  $H_q$ ,  $q > 4$ .

In the same section, we investigate single-nucleotide polymorphisms (SNPs) of some genes. In the case of SNP markers involving 3 bases, the fit of the card seq to that of the free group  $F_2$  is obtained, or not. The fit of the card seq of the selected SNP to that of  $F_2$  is well correlated to a lower risk of disease.

In Section 3.4, we analyze the binding domains and the card seq associated with motifs of the immediately early genes Fos, EGR1 and Myc. In such cases, the card seq of the group  $f_p$ , taken with the relation as the motif, is that of a free group  $F_r$  (in the finite range of indices). Most of the time, the motif of a transcription factor for a gene leads to the card seq of a  $F_r$ . This statement follows from an almost exhaustive investigation of transcription factors found in the databases defined in References [4,6].

However, it is also important to investigate the transcription factors with a group structure away from a free group. This is done in Section 3.5 with the claim that the lack of syntactical freedom (i.e., that the card seq of the gene is not that of a free group) is a marker of potential dysfunction of the gene through mutations or isoforms.

In Section 4, we show that group theoretical freedom correlates to the aperiodicity of motifs when the latter are seen as substitution sequences.

In the Conclusions, we offer some roads of progress in the connection of group theory to genetics.

### 3. Results

#### 3.1. The TATA Box, the Hecke Groups and More

The TATA box (also called the Goldberg–Hogness box) is a DNA sequence located in the core promoter region of genes in many eukaryotes, as well in archaea [11]. The TATA box is a non-coding sequence whose name comes from the fact that it contains a consensus sequence with repeating T and A base pairs [12,13]. The TATA box is a component of eukaryotic promoters in which it initiates the transcription of TATA-containing genes. The TATA box binds to the TATA-binding protein (TBP) and some other transcription factors. TBP binds to the minor groove of the TATA box via a region of antiparallel  $\beta$  sheets in the protein.

The regulation of gene transcription by transcription factors depends on the gene and is governed by the RNA polymerase II (PolII) transcription complex. In the core promoter of a typical PolII, they are key elements such as a TATA box.

Mutations such as insertions, deletions and point mutations to this consensus sequence can result in phenotypic changes. These changes can then be related to diseases such as gastric cancer, blindness, immunosuppression, Gilbert’s syndrome, etc. [12].

#### 3.2. Gilbert’s Syndrome

Gilbert’s syndrome is a genetic polymorphism associated with the gene UGT1A1, a phase II drug-metabolizing enzyme, which is essential in the metabolism of bilirubin and other drugs. The core promoter in UGT1A1 contains a TATA box located at position  $-28$  with respect to the transcriptional start site [14]. A polymorphism with  $AT(TA)_lTAA$  (with  $l = 5 \dots 8$ ) is common in all ethnic populations with  $l = 6$  as the major allele. Minor alleles with  $l > 6$  have less UGT1A1 transcription efficiency, leading to Gilbert’s syndrome, neonatal jaundice and toxicity in cancer chemotherapy [14].

In Table 2, we look at the finitely generated groups  $fp = \langle A, T | rel \rangle$ , where rel is the consensus sequence in the TATA box. The first two rows are for a standard TATA box. For this case, the group  $fp$  is found to have the same cardinality structure of cc of subgroups as the group  $H_3$  (the modular group) or the Hecke group  $H_4$ . Rows 3 and 4 are for the TATA box in the core promoter of the UGT1A1 gene for normal subjects, while rows 7 and 8 are for subjects with a Gilbert’s syndrome. In the former case, the group  $fp$  has a cardinality structure of cc of subgroups corresponding to the Hecke groups  $H_6$  and  $H_7$ , while in the latter case, the cardinality structure of cc of subgroups fits that of the Hecke groups  $H_4$  and  $H_3$  only up to index 8. Thus, we find that Gilbert’s syndrome is associated with an imperfect fit of the group  $G$  to a Hecke group.

**Table 2.** Group structure of a TATA box. Column 1 is for the selected consensus sequence (rows 4 to 6 are for the TATA box in the core promoter of UGT1A1 gene). Column 2 is for the cardinality sequence (card seq) of conjugacy classes (cc) of subgroups in the finitely generated group whose relation (rel) is the consensus sequence (cons seq). Column 3 identifies the Hecke group  $H_q = \langle A, T | A^2T^{-q} \rangle$ , which is close to the group under consideration (based on its card seq of subgroups). Column 4 refers to some references in the literature. Bold digits feature the fit to a Hecke group.

Rel: Cons Seq	Card Struct of cc of Subgroups	Group	Literature
TATAAAA	[1, 1, 2, 3, 2, 8, 7, 10, 18, 28, ...]	$H_3$	[6] (MA0108.1)
TATAAAA	[1, 3, 2, 8, 6, 19, 16, 69, 83, 238, ...]	$H_4$	[13]
$A(TA)_5TAA$	[1, 3, 3, 7, 6, 34, 42, 123, 319, 706, ...]	$H_6$	[14]
$A(TA)_6TAA$	[1, 1, 1, 1, 1, 34, 77, 79, 51, ...]	$H_7$	.
$A(TA)_7TAA$	[1, 3, 2, 8, 6, 19, 16, 171, 315, 1022, ...]	$\approx H_4$	.
$A(TA)_8TAA$	[1, 1, 2, 3, 2, 8, 7, 10, 308, 792, ...]	$\approx H_3$	.

### 3.3. Single Nucleotide Polymorphism

The canonical form of TBP-binding sites, the TATA box, is the best-studied regulatory element among human gene promoters. Tables identifying single-nucleotide polymorphisms (SNPs) in the (gene-dependent) TATA box have been collected in Reference [15].

At present, there are approximately  $10^8$  stored SNP markers that have been identified in the human genome and approximately  $10^{10}$  potentially possible markers. Most of them are neutral and do not affect health in any way. Markers in protein-coding regions of genes may damage proteins but are uncorrectable by treatment or lifestyle changes. However, a large number of the variants that have been identified are located in non-protein-coding regions and are presumed to affect gene expression regulation [16]. Regulatory SNPs in the TATA regions have biomedical usefulness and are correctable by medication and/or lifestyle. Ref. [15] collects 126 known SNP markers in 7 tables. We made use of these tables to compute the finitely generated group  $fp$  whose relation (rel) is the marker; see Table 3. For simplicity, we only took SNP markers built from 3 bases (and the exceptional SNP marker with 2 bases). We made the cardinality sequence of cc of subgroups (card seq) explicit. The computed closeness of the finitely generated group to the free group  $F_2$  correlates to a lower risk of illness. On the contrary, markers leading to a card seq away from that of  $F_2$  indicate a potential higher risk of illness. The symbol \* corresponds to the only two-base SNP marker in the table. In this case, the card seq is the same as the sequence for the fundamental group of 3-manifold  $m_{002} = net02_{00000}$ . The latter manifold is the smallest volume closed 3-manifold and is non orientable [17].

As a way of example, we take SNP markers in the first section of Table 3 that correspond to potential tumors in reproductive organs. Five of them show a card seq away from that of the free group  $F_2$ , and they also correspond to a potential higher risk of disease. The last two markers in the same section, whose card seq is close to that of  $F_2$ , are expected to produce a lower risk of breast cancer. Similar conclusions are valid for the SNP markers in other sections of Table 3.

**Table 3.** Group analysis of a few known and candidate SNP markers (taken from [15]) Column 1 is for the selected gene. Column 2 is for the SNP marker. Column 3 is for the card seq for the finitely generated group  $fp$  whose relation (rel) is the marker. Column 4 is for the reference paper and the letter indicates the heuristic confidence level of the candidate SNP marker (in alphabetical order from the best (A) to the worst (E)). The computed closeness of the finitely generated group to the free group  $F_2$ , most of time, correlates to a lower risk of illness, as described in [15]. The symbol \* corresponds to the only two-base SNP marker in the table. The card seq is the same as the sequence for the fundamental group of 3-manifold  $m_{002}$ . The latter manifold is the smallest volume closed 3-manifold and is non-orientable [17].

Gene	Rel: Marker	Card Seq of cc of Subgroups	Literature
ESR2	TTAAAAGGAA	[1, 7, 17, 114, 423, 4526, 30364, 293306 . . . ]	Table 1 in [15], B
HSD17B1	AGCCCAGAGC	[1, 3, 7, 26, 217, 124, 18443, 219870 . . . ]	., A
.	CAAGCCCAGA	[1, 7, 14, 109, 396, 3347, 19758, 188940 . . . ]	., A
PGR	AAAGGAGCCG	[1, 7, 17, 142, 475, 4125, 23509, 225871 . . . ]	., A
GSTM3	GGGTATAAAG	[1, 7, 14, 109, 396, 3347, 19758, 188940 . . . ]	., E
.	CCCCTCCCGC	[1, 3, 7, 26, 97, 624, 4163, 34470 . . . ]	., C
.	CCCTCCCGCT	.	., C
IL1B	AAAACAGCGA	[1, 7, 14, 89, 224, 1842, 10191, 86701 . . . ]	Table 2 in [15], A
CYP2A6	AAAGGCAAC	[1, 7, 17, 134, 683, 7077, 64225 . . . ]	., A
DHFR	GGGACGAGGG	[1, 3, 7, 26, 97, 624, 4163, 34470 . . . ]	., A
.	GGACGAGGGG	.	., A



**Table 3.** *Cont.*

Gene	Rel: Marker	Card Seq of cc of Subgroups	Literature
LEP	GGGGCGGGA	[1, 3, 7, 26, 97, 624, 4163, 34470 ...]	Table 3 in [15], C
GCG	TGCGCCTTGG	[1, 3, 7, 26, 119, 816, 4865, 40489 ...]	., B
GH1	TATAAAAAGG	[1, 7, 14, 109, 396, 3347, 19758, 188940]	., E
.	GTATAAAAAG	.	., D
.	GGTATAAAAA	.	., E
.	AGGGCCCCACA	[1, 3, 7, 26, 127, 860, 5661, 45710 ...]	., A
.	AAAGGGCCCC	[1, 3, 10, 67, 266, 3458, 30653, 312237 ...]	., A
.	AAAGGGCCA	.	., A
NOS2	TCTTGGCTGC	[1, 3, 7, 26, 97, 624, 4163, 34470 ...]	Table 4 in [15], A
TPI1	ATATAAGTGG	[1, 3, 7, 30, 125, 856, 4832, 40246 ...]	., B
GJA5	TATTAAACAC	[1, 3, 10, 35, 140, 921, 5778, 47238 ...]	., E
HBD	AAAAGGCAGG	[1, 3, 7, 26, 97, 624, 4163, 34470 ...]	Table 5 in [15], A
F2	AACCCAGAGG	[1, 3, 7, 26, 127, 860, 5661, 45710 ...]	., A
F8	GGAAGAGGGA	[1, 3, 2, 7, 4, 18, 9, 27, 36, 68 ...] *	Table 6 in [15], A
F3	GCGCGGGGCA	[1, 3, 7, 26, 97, 624, 4163, 34470 ...]	., A
F11	TTTTIAGTAA	.	., D
.	TTTTIAGTAA	[1, 7, 17, 114, 423, 4526, 30364, 293306 ...]	., A
.	AAGGAAATTT	[1, 3, 7, 26, 195, 1692, 11803, 73192 ...]	., A
AR	GTGGAAGATT	[1, 3, 7, 34, 139, 931, 5208, 43867 ...]	Table 7 in [15], A
.	CCACGACCCG	[1, 7, 20, 167, 754, 7232, 60860, 683597 ...]	., D
MTHFR	TCCCTCCA	[1, 3, 7, 26, 97, 624, 4163, 34470 ...]	., A
DMNT1	TGTGTGGCCCG	.	., A
.	GTGTGTGCC	.	., A
.	GACGAGCCCA	[1, 3, 7, 42, 131, 912, 6011, 47322 ...]	., A
NR5A1	ACAAGAGAAA	[1, 3, 7, 26, 97, 624, 4163, 34470 ...]	., A
.	GGTGTGAGAG	[1, 7, 14, 89, 264, 1987, 11086, 93086 ...]	., A

### 3.4. A Few DNA/Protein Complexes and Their Transcription Factors

As mentioned in the Introduction, most transcription factors have a binding domain whose finitely generated group  $fp$  has a subgroup signature equal to that of a free group  $F_r$  of rank  $r$ , with  $r \leq 3$  corresponding to  $r + 1 \leq 4$  distinct letters. An almost exhaustive search has been performed by using the catalogs found in [4,6]. We first give details of our calculations for a few immediate early genes. Then, in Section 3.5, most found counterexamples are summarized.

#### 3.4.1. Immediate Early Genes and Their Motifs

Immediate early genes (IEGs) are genes that are activated transiently and rapidly in response to a wide variety of cellular stimuli. They represent a standing response mechanism that is activated at the transcription level in the first round of response to stimuli, before any new proteins are synthesized. The earliest known and best characterized include *c-fos*, *c-myc* and *c-jun*, genes that were found to be homologous to retroviral oncogenes. IEGs are well known as early regulators of cell growth and differentiation signals. However, other findings suggest roles for IEGs in many other cellular processes as ‘gateways to genomic response’. Many IEG products are natural transcription factors or other DNA-binding proteins. Important classes of IEG products include secreted proteins, cytoskeletal proteins and receptor subunits.

Some IEGs such as ZNF268 and Arc have been implicated in learning, memory and long-term potentiation. Neuronal IEGs are used prevalently as a marker to track brain activities in the context of memory formation and the development of psychiatric disorders [18].

The group structure of the motifs of some IEGs in the Fos, EGR and Myc classes is summarized in Table 4.

**Table 4.** Group structure of motifs for transcription factors of immediately early genes Fos, EGR and Myc. Most of the time, the card seq of the group defined with the relation/motif is the free group  $F_2$  (for a 3 letter motif) or  $F_3$  (for a 4 letter motif). There are two exceptions for the EGR1 gene, depending on the selected motif, where the card seq corresponds to the modular group  $H_3$  or the Baumslag–Solitar group  $BS(-1, 1)$ , which is the fundamental group of the Klein bottle. The card seq for  $H_3$  is in Table 2. The card seq for  $BS(-1, 1)$  is  $[1, 3, 2, 5, 2, 7, 2, 8, 3, 8, 2, 13, 2, 9, 4, \dots]$ .

Gene	Rel: Motif	Card Seq	Literature
Fos	TGAGTCA	$F_3$	[19]
.	TGACTCA	$F_3$	[6], MA MA0099.2
EGR1	GCGTGGGCG	$F_2$	[6], MA0162.1
.	CCGCCCCCG	$H_3$	., MA0162.2
.	CCGCCCCCGC	$BS(-1, 1)$	., .
.	ACGCCCACGCA	$F_2$	., MA0162.3
.	GGCCCACGC	.	., MA0162.4
EGR2	CCGCCACGC	.	., MA0472.1
.	ACGCCCACGCA	.	., MA0472.2
EGR3,EGR4	ACGCCCACGCA	.	., [ MA0732.1, MA0733.1]
Myc	CACGTG	$F_3$	[19]
.	CGCACGTGGT	.	[6], MA0147.1
.	CCCACGTGCTT	.	., MA0147.2
.	CCACGTGC	.	., MA0147.3
Mycn, Max::Myc, etc	GACCACGTGGT, etc.	.	., [MA0104.1, etc.]

### 3.4.2. The DNA-Binding Domain Fos

The Fos family (as well as the Jun family) are eukaryotic transcription factors that heterodimerize to form complexes binding elements such as 5'-TGAGTCA-3' DNA elements [19]. The X-ray crystal structure was determined and the bZIP region (with 62 aa) of the c-Fos protein bound to DNA is available in the protein data bank as PDB: 1FOS. The protein secondary structure of this subunit of c-Fos protein consists of two alpha helices, as shown in Figure 1.

Let us consider the group  $fp = \langle A,T,G,C|rel \rangle$  on 4 letters with the relation  $rel = bind = TGAGTCA$ . The card seq of  $fp$  up to index 6 is that of the free group  $F_3 = \langle A,T,G,C|AGTC \rangle$  of rank 3. One can use the coset enumeration (with the Todd–Coxeter procedure) to check that, up to index 6, the permutation groups organizing the cosets in the cc of groups  $fp$  and  $F_3$  are the same. This shows that both groups are close, at least in the finite range of subgroups. However,  $fp$  and  $F_3$  are not the same group. Incidentally, the group  $fp' = \langle A,T,G,C|AGTC, bind \rangle$ , with the joint relations of  $fp$  and  $F_3$ , is close to the free group  $F_2 = \langle x, y, z|xyz \rangle$  on two generators in the sense that the cardinality sequence of the cc of subgroups is that of  $F_2$ , up to the higher index 9 that we could reach in our calculations.

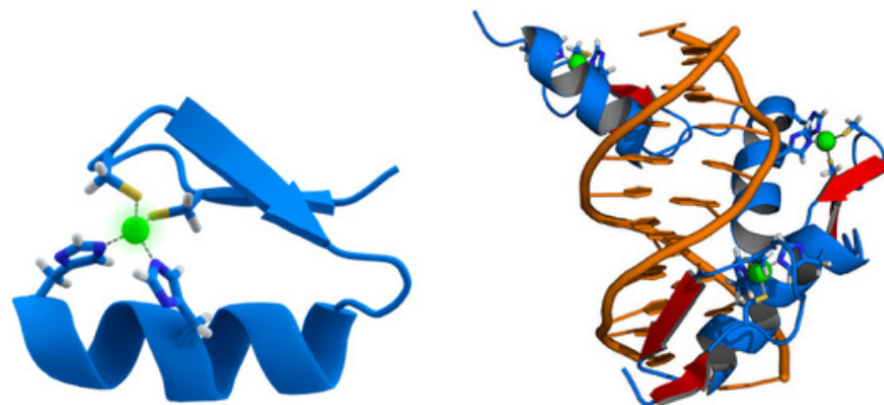
Similarly, the finitely generated groups  $fp = \langle A,T,G,C|rel \rangle$ , where the relation is with the whole DNA chains  $rel = AATGGATGAGTCATAGGAGA$  (1FOS\_1) or  $rel = TTCTCCTATGACTCATCCAT$  (1FOS\_2) involved in the DNA/protein Fos complex (PDB:1FOS), have the same card seq as  $F_3$  up to index 6.



**Figure 1.** The DNA-binding domain of the immediate early gene Fos. The name in the protein data bank is 1FOS.

### 3.4.3. The DNA-Binding Domain EGR1

The DNA-binding domain EGR1 (for early growth response protein 1) is a mammalian transcription factor also called ZNF268 (the zinc finger protein 268). This is because the protein encoded by the EGR1 gene has the Cys<sub>2</sub>His<sub>2</sub>-like fold structure of a zinc finger, as shown in Figure 2. It binds to the motif 5'-bind-3' [20], with bind = GCG(T/G)GGCG.



**Figure 2.** (Left) Cartoon representation of the Cys<sub>2</sub>His<sub>2</sub> zinc finger motif, consisting of an  $\alpha$ -helix and an antiparallel  $\beta$ -sheet. The zinc ion (green) is coordinated by two histidine residues and two cysteine residues. (Right) Cartoon representation of the protein ZNF268 (blue) containing three zinc fingers in complex with DNA (orange). The coordinating amino acid residues and zinc ions (green) are highlighted. The name of the DNA-binding domain in the protein data bank is 4R2A.

The protein in the DNA-binding domain EGR1 is a nuclear protein and functions as a transcriptional regulator. The products of the target genes that it activates are required for differentiation and mitogenesis. When located in the brain, it has an essential role in memory formation and in brain neuron epigenetic reprogramming. An X-ray crystal structure is available in the protein data bank as PDB: 4R2A. In such a EGR1 DNA-binding domain, the DNA chains are rel = AGCGTGGGCGT and rel = TACGCCACGC.



As for the Fos domain above, let us consider the group  $fp = \langle A,T,G,C|rel \rangle$  on 4 letters with the relation bind or rel. The card seq of  $fp$  up to index 6 is similar to that of the free group  $F_3 = \langle A,T,G,C|ATGC \rangle$  of rank 3. One can use the coset enumeration (with the Todd–Coxeter procedure) to check that, up to index 6, the permutation groups organizing the cosets in the cc of subgroups of  $fp$  and  $F_3$  are the same. This shows that both groups are close, at least in the finite range of subgroups. However,  $fp$  and  $F_3$  are not the same group. Again, the groups built from the joint relations of  $fp$  and  $F_3$  are of rank 2, but the cardinality structure of cc of subgroups is not that of  $F_2$ .

The group  $fp' = \langle A,T,G,C|ATGC, bind \rangle$ , with the joint relations of  $fp$  and  $F_3$ , is close to the free group  $F_2 = \langle x, y, z | xyz \rangle$  on two generators in the sense that the card seq is that of  $F_2$ , up to the higher index 9 that we could reach in our calculations.

The early growth response protein 1 contains the chain of amino acids

GPLGS ERPYACPVESCDRRFRSDELTRHIRIHTG QKPFQCRICMRNFSRSDHLT-  
THIRHTG EKPFACDICGRKFARSDEKRKHTKIHLR QKD.

The central portion of the protein contains  $86 = 30 + 28 + 28$  aa decomposed into 3 zinc fingers with the following secondary structure (letter H is for the  $\alpha$ -helix segment, letter E is for the  $\beta$ -sheet segment and letter C is for the random coil segment)

CCCECCCCCCCCCEECHHHHHHHHHHHHHHHH CCCECCCCCCCCCEECHHHHHH-  
HHHHHHHHH CCCECCCCCCCCCEECHHHHHHHHHHHHHHHH C

Taking the former 3-letter chain as the relation of a finitely generated group on 3 letters (and rank 2), we get the cardinality sequence for the cc of its subgroups as  $[1, 3, 7, 26, 112, 717, \dots]$ , which fits the cardinality sequence of cc of subgroups of the free group  $F_2$  only up to the index 4.

#### 3.4.4. The DNA-Binding Domain Myc

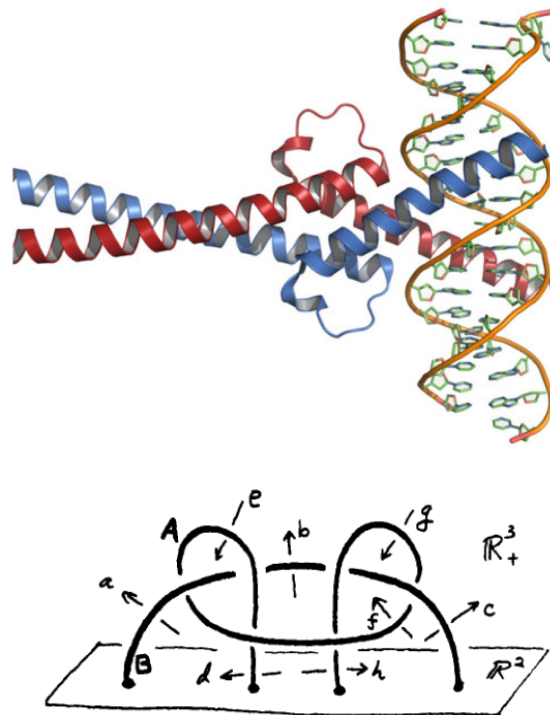
Myc proto-oncogene is a transcription factor encoding a nuclear phosphoprotein that plays a role in cell cycle progression, apoptosis and cellular transformation [21]. The protein contains a basic helix-loop-helix zipper (bHLHZ) structural motif. The encoded protein forms a heterodimer with the related transcription factor Max, as shown in Figure 3. Amplification of this gene is frequently observed in numerous human cancers. Translocations involving this gene are associated with Burkitt lymphoma and multiple myeloma in human patients.

The bHLHZ domain of Myc-Max binds to the common DNA (palindromic) target  $5'$ -CACGTG- $3'$ . In the protein data bank, the reference of the complex is PDB: 1NKP. The whole DNA chain is

rel = CGAGTAGCACGTGCTACTC (1NKP\_1).

Let us consider the group  $fp = \langle A,T,G,C|rel \rangle$  on 4 letters with the relation rel = bind = CACGTG. The conjugacy classes of  $fp$  up to index 6 have card seq equal to that of the free group  $F_3 = \langle A,T,G,C|GACT \rangle$  of rank 3. Again, one can use the coset enumeration (with the Todd–Coxeter procedure) to check that, up to index 6, the permutation groups organizing the cosets in the cc of groups  $fp$  and  $F_3$  are the same. This shows that both groups are close, at least in the finite range of subgroups. However,  $fp$  and  $F_3$  are not the same group. The group  $fp' = \langle A,T,G,C|GACT, bind \rangle$ , with the joint relations of  $fp$  and  $F_3$ , is close to the group  $\pi_2$  defined in Figure 3 (Down). The group  $\pi_2 = \langle x, y, z | (x, (y, z)) = z \rangle$  is the fundamental group of the union of two links  $A$  and  $B$ , which are not splittable. The proof is in Refs. [22] and ([23] p. 90) and follows from the fact that  $\pi_2$  is not a free group. The group  $fp'$  is close to  $\pi_2$  in the sense that the cardinality sequence  $[1, 3, 10, 51, 164, 1365, 9422, 81594, 721305, \dots]$  of the cc of subgroups is that of  $\pi_2$ , up to the higher index 9 that we could reach in our calculations.

The non-closeness of  $fp'$  to  $F_2$  and the fact that  $\pi_2$  is not free are distinguishing features of the Myc domain. It is tempting to associate such features with a potential abnormal replication.



**Figure 3.** (Up) Crystal structure of Myc and Max in complex with DNA. (Down) The link  $L = A \cup B$  (which is supposed to control the binding domain Myc) is attached to the plane  $R^2$  in the half-space  $R^3_+$ . It is not splittable. This can be proven by checking that the fundamental group  $\pi = \pi_2(L)$  is not free [22] and ([23] p. 90). One gets  $\pi_2 = \langle x, y, z | (x, (y, z)) = z \rangle$ , where  $(..)$  means the group theoretical commutator. The cardinality sequence of cc of subgroups of  $\pi_2$  is  $[1, 3, 10, 51, 164, 1365, 9422, 81594, 721305, \dots]$ .

### 3.5. Genes Whose Transcription Factors Have a Group Structure Away from a Free Group

We analyzed the group structure of motifs for some transcription factors that do not lead to free groups. This is shown in Table 5. A short account of the function or dysfunction of the corresponding genes is given in Table 6. It is observed that several transcription factors whose group structure is away from a free group have the same card seq. We conjecture that it is indicative of a related 3-dimensional structure of the corresponding domain, although these families do not fit the standard classification [6].

**Table 5.** Group structure of motifs for some transcription factors that are not leading to free groups. The card seq for  $\pi_1$  is  $[1, 4, 1, 2, 4, 2, 1, 7, 2, 2, 4, 2, 2, 8, 1, 2, 7, 2, 3, \dots]$ ; for  $\pi'_1$  it is  $[1, 1, 1, 2, 1, 3, 3, 1, 2, 2, 1, 1, 9, 2, 14, 2, 1, \dots]$ . The card seq for  $\pi_2$  is already in Figure 3 as  $[1, 3, 10, 51, 164, 1365, 9422, 81594, 721305, \dots]$ . The card seq for  $\pi_3$  is  $[1, 7, 14, 89, 264, 1987, 11086, 93086, \dots]$ ; for  $\pi'_3$ , it is  $[1, 7, 50, 867, 15906, 570528, \dots]$ ; for  $\pi''_3$ , it is  $[1, 7, 50, 739, 15234, 548439, \dots]$ ; for  $\pi_3^{(3)}$ , it is  $[1, 7, 41, 668, 14969, 550675]$ . The card seq for  $\pi_4$  is  $[15, 82, 1583, 30242, \dots]$ . The index  $i$  in  $\pi_i$  refers to the rank of the group under examination. The three sections are for motifs on 2, 3 and 4 letters, respectively.

Gene	Rel: Motif	Card Seq	Literature
NKX6-2	TAATTAA	$H_3$	[6], [MA0675.1, MA0675.2]
HoxA1, HoxA2 POU6F1, Vax RUNX1	TAATTA . TGTGGT	$\pi_1$ . .	[6], [MA1495.1, MA0900.1] ., [MA0628.1, MA0722.1] ., MA0511.1

**Table 5.** Cont.

Gene	Rel: Motif	Card Seq	Literature
RUNX1 EHF	TGTGGTT CCTTCCTC	$\pi_1'$ .	[6], MA0002.2 ., MA0598.1
POU6F1 PITX2 ELK4	TAATGAG TAATCCC CTTCCGG	$\pi_2$ . .	[6] MA1549.1 ., [MA1547.1, MA1547.2] ., MA0076.2
OTX2, Dmbx1 PitX1, PitX2, PitX3, OTX1 N-box	GGATTA TAATCC TTCCGG	$\pi_3$ . .	[6], [MA0712.2, MA0883.1] ., [MA0682.1, MA0711.1] [24]
p53 GZF1 NF-kappa-B STAT1 . STAT4	CACATGTCCA TGCGCGTCTATA GGGAATTTCC TTCCCCGGAA TTCCAGGAA TTCCAGGAAA	$\pi_3'$ . . . . .	[25] [4] [6], [MA0107.1, MA1911.1] ., MA0137.2 ., MA0137.3 ., MA0518.1
FOSL1::Jun USF2 PAX1 STAT2	ATGACGTCAT GTCATGTGACC CGTCACGCATGA TTCCAGGAAG	$\pi_3''$ . . .	[6], MA1129.1 ., MA0626.1 ., MA0779.1 ., MA0144.1
FOS MAFA, MAFF,MAFK	GATGACGTCATCA TGCTGAGTCAGCA	$\pi_3^{(3)}$ .	[6], MA1951.1 ., [MA1521.1, MA0495.2, MA0946.2]
CREB USF2 SMAD3, SMAD5	TGACGTCA GGTCACGTGACC GTCTAGAC	$\pi_4$ . .	[6], [MA0018.2, MA018.3] ., MA0526.4 ., [MA0795.1, MA1557.1], [26]

The DNA-Binding Domain of p53

Tumor protein p53 (also called tumor suppressor p53) has been called the guardian of the genome. The main reason behind this status is the critical role that p53 plays in preventing cancer development. p53's role in tumor suppression is due to its ability to induce the apoptosis, cell cycle arrest and senescence of pre-cancerous cells. However, it also regulates other genes involved in metabolism.

According to [25], a motif for p53 is the DNA sequence CACATGTCCA. In our Table 5, the attached card seq in the finite range of indices is that of a group  $\pi_3'$ . However, there are motifs leading to a card seq associated with the free group  $F_3$  or with other non-free groups that are not of type  $\pi_3'$ . This may be due to the fact that p53 has many isoforms to fill its role.

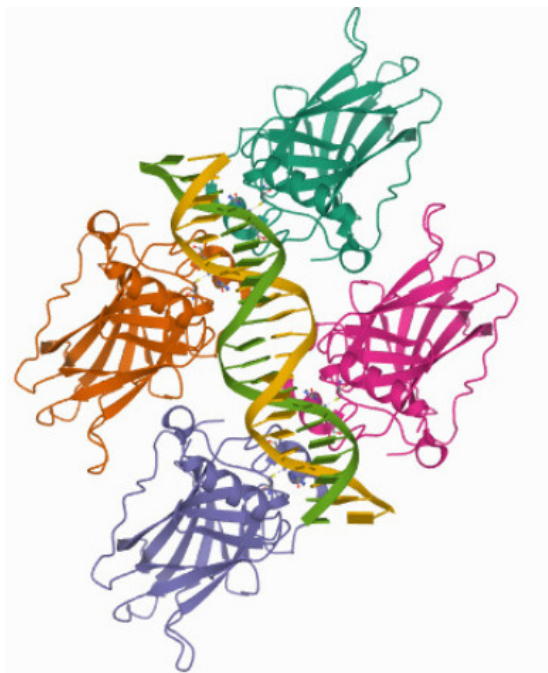
In Figure 4, we draw the crystal structure of the p53 domain for the binding domain of the PDB sequence 4HJE [27]. The p53 domain forms a tetramer, but other symmetries of the binding domain of p53 may be found.

**Table 6.** A short account of the function or dysfunction (through mutations or isoforms) of genes associated with transcription factors and sections in Table 5.

Gene	Type	Function	Dysfunction
NKX6-2	homeobox	central nervous system, pancreas	spastic ataxia
HoxA1	homeobox	embryonic devt of face and heart	autism
HoxA2	.	.	cleft palate

Table 6. Cont.

Gene	Type	Function	Dysfunction
Pou6F1 Vax	.	neuroendocrine system forebrain development	clear cell adenocarcinoma craniofacial malform.
RunX1	Runt-related	cell differentiation, pain neurons	myeloid leukemia
EHF	homeobox	epithelial expression	carcinogenesis, asthma
PitX2 ELK4	. Ets-related	eye, tooth, abdominal organs serum response for c-Fos	Axenfeld–Rieger syndrome
OTX1,OTX2 Dmbx1	homeobox .	brain and sensory organ devt .	medulloblastomas farsightedness and strabismus
PitX1	.	organ devt, left–right asymmetry	autism, club foot
PitX3 N-box	. Ets-related	lens formation in eye synaptic expression	congenital cataracts drug sensitivity
p53 GZF1 NF-kappa-B	p53 domain Zinc fingers .	‘Guardian of the genome’ protein coding DNA transcription, cytokines	cancers short stature, myopia apoptosis
STAT1	Stat family	signal activator of transcription	immunodeficiency 31
STAT4	Stat family	signal activator of transcription	rheumatoid arthritis
FOSL1::Jun USF2 PAX1	leucine zipper helix-loop-helix paired box	cellular proliferation transcription activator fetal development	marker of cancer Klippel–Feil syndrome
FOS Maf	leucine zipper .	cellular proliferation pancreatic development	cancers congenital cerulean cataract
CREB USF2 SMAD	bZIP helix-loop-helix homeo domain	neuronal plasticity transcription activator cell development and growth	Alzheimer’s disease Alzheimer’s disease



**Figure 4.** Crystal structure of p53 binding domain. The reference number in the protein data bank is 4HJE.

#### 4. Discussion

According to Reference [1], aperiodicity is the correlate of syntactical freedom of ordering rules. How can we check this statement in the realm of transcription factors?

First, we introduce the concept of a general substitution rule in the context of free groups. A general substitution rule  $\rho$  on a finite alphabet  $\mathcal{A}_r$  on  $r$  letters is an endomorphism of the corresponding free group  $F_r$  [28] (Definition 4.1). The endomorphism property means the two relations  $\rho(uv) = \rho(u)\rho(v)$  and  $\rho(u^{-1}) = \rho^{-1}(u)$ , for any  $u, v \in F_r$ .

A special role is played by the subgroup  $\text{Aut}(F_r)$  of automorphisms of  $F_r$ . We introduce the map  $\alpha : F_r \rightarrow \mathbb{Z}_r$  from  $F_r$  to the Abelian group  $\mathbb{Z}_r$  in order to investigate the substitution rule  $\rho$  with the tools of matrix algebra.

The map  $\alpha$  induces a homomorphism  $M : \text{End}(F_r) \rightarrow \text{Mat}(r, \mathbb{Z})$ . Under  $M$ ,  $\text{Aut}(F_r)$  maps to the general linear group of matrices with integer entries  $GL(r, \mathbb{Z})$ . Given  $\rho$ , there is a unique mapping  $M(\rho)$  that makes the map diagram commutative [28] (p. 68). The substitution matrix  $M(\rho)$  of  $\rho$  may be specified by its elements at row  $i$  and column  $j$  as follows

$$(M(\rho))_{i,j} = \text{card}(\rho_{a_i}(a_j)).$$

We will apply this approach to binding motifs of transcription factors. The binding motif  $\text{rel}$  in the finitely presented group  $fp = \langle A, T, G, C | \text{rel}(A, T, G, C) \rangle$  is split into appropriate segments so that  $\text{rel} = \text{rel}_A \text{rel}_T \text{rel}_G \text{rel}_C$  with the substitution rules  $A \rightarrow \text{rel}_A, T \rightarrow \text{rel}_T, G \rightarrow \text{rel}_G, C \rightarrow \text{rel}_C$ .

We are interested in the sequence of finitely generated groups

$$f_p^{(l)} = \langle A, T, G, C | \text{rel}(\text{rel}(\text{rel} \cdots (A, T, G, C))) \rangle \quad (\text{with rel applied } l \text{ times})$$

whose card seq is the same at each step  $l$  and equal to the card seq of the free group  $F_r$  (in the finite range of indices that it is possible to check with the computer).

Under these conditions, we will see that (group) syntactical freedom correlates to the aperiodicity of sequences.

##### 4.1. Aperiodicity of Substitutions

There is no definitive classification of aperiodic order, the intermediate between crystalline order and strong disorder, but in the context of substitution rules, some criteria can be found. We need a few definitions.

A non-negative matrix  $M \in \text{Mat}(d, \mathbb{R})$  is one whose entries are non-negative numbers. A positive matrix  $M$  (denoted  $M > 0$ ) has at least one positive entry. A strictly positive matrix (denoted  $M \gg 0$ ) has all positive entries. An irreducible matrix  $M = (M_{ij})_{1 \leq i, j \leq d}$  is one for which there exists a non-negative integer  $k$  with  $(M^k)_{ij} > 0$  for each pair  $(i, j)$ . A primitive matrix  $M$  is one such that  $M^k$  is a strictly positive matrix for some  $k$ .

A Perron–Frobenius (PF for short) eigenvector  $v$  of an irreducible non-negative matrix is the only one whose entries are positive:  $v > 0$ . The corresponding eigenvalue is called the PF eigenvalue.

We will use the following criterion [28] (Corollary 4.3). A primitive substitution rule  $\rho$  of substitution matrix  $M(\rho)$  with an irrational PF-eigenvalue is aperiodic.

A well-studied primitive substitution rule is the Fibonacci rule  $\rho = \rho_F : a \rightarrow ab, b \rightarrow a$  of substitution matrix  $M_F = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$  and PF eigenvalue equal to the golden ratio  $\lambda_{PF} = \tau = (\sqrt{5} + 1)/2$  [28] (Example 4.6). As expected, the irrationality of  $\lambda$  corresponds to the aperiodicity of the Fibonacci sequence.

The sequence of Fibonacci words is as follows:

$$a, b, ab, aba, abaab, abaababa, abaababaabaab, \dots$$

The words have lengths equal to the Fibonacci numbers 1, 1, 2, 3, 5, 8, 13, 21,  $\dots$



It is straightforward to check that all finitely generated groups  $f_p^{(l)}$  whose relations  $\text{rel}(a, b) = ab, aba, abaab, abaababa, \dots$  have a card seq whose elements are 1's, as for the card seq of the free group  $F_1$ . The Fibonacci sequence is our first example where group syntactical freedom correlates to aperiodicity.

We will now provide examples taken for transcription factors involving 2, 3 or 4 letters.

#### 4.1.1. A Two-Letter Sequence for the Transcription Factor of Gene DBX in Drosophila Melanogaster

Let us consider the motif  $\text{rel} = \text{TTTATTA}$  for the gene DBX in drosophila melanogaster (fruit flies) [6] (MA0174.1). The roles of the DBX gene include neuronal specification and differentiation.

We split  $\text{rel}$  into two segments so that  $\text{rel} = \text{rel}_A \text{rel}_T$  with the substitution maps  $A \rightarrow \text{rel}_A = \text{T T T A}$ ,  $T \rightarrow \text{rel}_T = \text{T T A}$  to produce the substitution sequence

$$A, T, AT, \text{TTTATTA}, \text{TTATTATTATTATTATTATTATTTA} \dots$$

The substitution matrix for this sequence is  $M = \begin{pmatrix} 1 & 3 \\ 1 & 2 \end{pmatrix}$ ; it is a primitive matrix of PF eigenvalue  $\lambda_{PF} = (3 + \sqrt{13})/2$  so that the sequence associated with the DBX factor is aperiodic.

Similarly to the Fibonacci generator rules, all finitely generated groups  $f_p^{(l)}$  whose relations are  $\text{rel}(A, T) = AT, \text{TTTATTA}, \text{TTATTATTATTATTATTATTATTTA} \dots$  have a card seq whose elements are 1's as for the card seq of the free group  $F_1$ .

The sequence for the DBX transcription factor is our second example where group syntactical freedom correlates to aperiodicity.

#### 4.1.2. A Three-Letter Sequence for the Transcription Factor of Gene EGR1

The transcription factor of gene EGR1 was investigated in Section 3.4.1. The selected motif is  $\text{rel} = \text{GCGTGGGCG}$ . We split  $\text{rel}$  into two segments so that  $\text{rel} = \text{rel}_C \text{rel}_G \text{rel}_T$  with the substitution maps  $C \rightarrow \text{rel}_C = G$ ,  $G \rightarrow \text{rel}_G = \text{CGT}$ ,  $T \rightarrow \text{rel}_T = \text{GGGCG}$  to produce the substitution sequence:

$$C, G, T, \text{CGT}, \text{GCGTGGGCG}, \text{CGTGCGTGGGCGGCGTCGTCGTGCGT} \dots$$

The substitution matrix for this sequence is  $M = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 4 \\ 0 & 1 & 0 \end{pmatrix}$ ; it is a primitive matrix

(since  $M^2 \gg 0$ ) whose eigenvalues follow from the vanishing of the polynomial  $-\lambda^3 + \lambda^2 + 5\lambda + 1 = 0$ . There are three real irrational roots  $\lambda_1 \approx 2.86619$ ,  $\lambda_2 \approx -0.21075$ ,  $\lambda_3 \approx -1.65544$ . The PF eigenvalue is  $\lambda_{PF} = \lambda_1$  with an eigenvector of (positive) entries  $(1, \lambda_1/(\lambda_1^2 - \lambda_1 - 4), 1/(\lambda_1^2 - \lambda_1 - 4))^T \approx (1, 2.12485, 0.74134)^T$ .

It follows that the selected sequence for the EGR1 gene is aperiodic.

All finitely generated groups  $f_p^{(l)}$  whose relations are  $\text{rel}(C, G, T) = \text{CGT}, \text{GCGTG GCG}, \text{CGTGCGTGGGCGGCGTCGTCGTGCGT} \dots$  have a card seq whose elements are those of the free group  $F_2$ .

The sequence for the EGR1 transcription factor is our third example where group syntactical freedom correlates to aperiodicity.

#### 4.1.3. A Four-Letter Sequence for the Transcription Factor of the Fos Gene

The transcription factor of gene Fos was investigated in Section 3.4.1. The selected motif is  $\text{rel} = \text{TGAGTCA}$ .

We split  $\text{rel}$  into two segments so that  $\text{rel} = \text{rel}_A \text{rel}_T \text{rel}_C \text{rel}_C$  with the substitution maps  $A \rightarrow \text{rel}_A = T$ ,  $T \rightarrow \text{rel}_T = G$ ,  $G \rightarrow \text{rel}_G = \text{AGTC}$ ,  $C \rightarrow \text{rel}_C = A$ , to produce the substitution sequence

$A, T, G, C, ATGC, TGAGTCA, GAGTCTAGTCGAT \dots$

The substitution matrix for this sequence is  $M = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$ . It is a primi-

tive matrix ( $M^4 \gg 0$ ) whose eigenvalues follow from the vanishing of the polynomial  $\lambda^4 - \lambda^3 - \lambda^2 - \lambda - 1$ . There are two real eigenvalues  $\lambda_1 \approx 1.92756$  and  $\lambda_2 \approx -0.77480$ , as well as two complex conjugate eigenvalues  $\lambda_{3,4} \approx -0.07637 \pm 0.81470i$ .

The PF eigenvalue is  $\lambda_{PF} = \lambda_1$  with an eigenvector of (positive) entries  $\approx (1, 0.37298, 0.40211, 0.20861)^T$ .

It follows that the selected sequence for the Fos gene is aperiodic.

All the finitely generated groups  $f_p^{(l)}$  whose relations are:

$$\text{rel}(A,C,G,T) = ATGC, TGAGTCA, GAGTCTAGTCGAT \dots$$

have a card seq whose elements are 1, 7, 41, 604, 13753, 504243  $\dots$ .

which is the card seq of the free group  $F_3$ .

The sequence for the Fos transcription factor is our fourth example where group syntactical freedom correlates to aperiodicity.

## 5. Conclusions

We made use of group theory for investigating transcription factors in genetics. Finite group theory plays a major role in the attempts to model the genetic code; see [29,30] and other references therein. Finitely generated groups (whose cardinality is infinite) are necessary to deal with the secondary structures of proteins [2]. It was already noted that many structures for the protein secondary codes tend to be close to free groups. The card seq for such codes is model-dependent. In the map from amino acids to proteins, the transcription factors play a critical role. The study of the group theoretical structure of TFs has been our goal in the present paper. The DNA motifs that serve as a relation for the corresponding  $fp$  groups are, in general, short sequences with around 10 amino acids. Taking random sequences instead of the gene-specific DNA sequences in TFs also leads to a majority of cases where the card seq of  $fp$  is close to a free group  $F_r$  and less frequent cases where the card seq is away. However, motifs in TFs are codes with a particular meaning—the specific gene function or dysfunction. In this sense, we found it appropriate to use the concept of ‘syntactical freedom’ to qualify most TFs and to associate the lack of syntactical freedom with a potential source of illness. In our context, syntactical freedom means free groups and aperiodicity.

Potentially, there are many potential applications of our group theoretical method of TFs for characterizing genes with problematic mutations, such as cancer and Alzheimer’s disease. More work will be performed in the future.

Another interesting line of research is about the neurogenetic correlation of human consciousness and the related TFs. If the reader is interested in studying this research further, we recommend [24,31–33].

As a final note, we refer to the paper [34,35] in the domain of quantum gravity, where the ordering rules with syntactical freedom are those of quasicrystals instead of those of the biological crystal structures.

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

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

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

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

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

## References

1. Irwin, K. The code-theoretic axiom; the third ontology. *Rep. Adv. Phys. Sci.* **2019**, *3*, 39. [CrossRef]
2. 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. [CrossRef]
3. Planat, M.; Aschheim, R.; Amaral, M.M.; Fang, F.; Irwin, K. Graph coverings for investigating non local structures in protein, music and poems. *Science* **2021**, *3*, 39. [CrossRef]
4. Lambert, S.A.; Jolma, A.; Campitelli, L.F.; Das, P.K.; Yin, Y.; Albu, M.; Chen, X.; Talpale, J.; Hughes, T.R.; Weirauch, M.T. The human transcription factors. *Cell* **2018**, *172*, 650–665. Available online: [http://www.edgar-wingender.de/huTF\\_classification.html](http://www.edgar-wingender.de/huTF_classification.html) (accessed on 1 September 2021).
5. Wingender, E.; Schoeps, T.; Dönitz, J. TFClass: An expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* **2013**, *T1*, D165–D170. [CrossRef] [PubMed]
6. Sandelin, A.; Alkema, W.; Engström, P.; Wasserman, W.W.; Lenhard, B. JASPAR: An open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **2004**, *32*, D91–D94. Available online: <https://jaspar.genereg.net/> (accessed on 1 September 2021). [CrossRef]
7. Planat, M.; Giorgetti, A.; Holweck, F. Saniga, M. Quantum contextual finite geometries from dessins d'enfants. *Int. J. Geom. Meth. Mod. Phys.* **2015**, *12*, 1550067. [CrossRef]
8. Hall, M., Jr. Subgroups of finite index in free groups. *Can. J. Math.* **1949**, *1*, 187–190. [CrossRef]
9. Kwak, J.H.; Nedela, R. Graphs and their coverings. *Lect. Notes Ser.* **2007**, *17*, 118.
10. The Modular Group. Available online: [https://en.wikipedia.org/wiki/Modular\\_group](https://en.wikipedia.org/wiki/Modular_group) (accessed on 1 October 2021).
11. Suzuki, Y.; Tsunoda, T.; Sese, J.; Taira, H.; Mizushima-Sugano, J.; Hata, H.; Ota, T.; Isogai, T.; Tanaka, T.; Nakamura, Y.; et al. Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.* **2001**, *11*, 677–684. [CrossRef]
12. TATA Box. Available online: [https://en.wikipedia.org/wiki/TATA\\_box](https://en.wikipedia.org/wiki/TATA_box) (accessed on 1 September 2021).
13. Wang, Y.; Jensen, R.C.; Stumph, W.E. Role of TATA box sequence and orientation in determining RNA polymerase II/III transcription specificity. *Nucleic Acids Res.* **1996**, *24*, 3100–3106. [CrossRef]
14. Li, Y.; Buckley, D.; Wang, S.; Klaassen, C.D.; Zhong, X.B. Phenobarbital-Responsive Enhancer Module of the UGT1A1. *Drug Metab. Disp.* **2009**, *37*, 1978–1986. [CrossRef] [PubMed]
15. Chadaeva, I.V.; Ponomarenko, P.M.; Rasskazov, D.A.; Sharypova, E.B.; Kashina, E.V.; Zhechev, D.A.; Drachkova, I.A.; Arkova, O.V.; Savinkova, L.K.; Ponomarenko, M.P.; et al. Candidate SNP markers of reproductive potential are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *BMC Genom.* **2018**, *19*, 16–38. [CrossRef] [PubMed]
16. Zerbino, D.R.; Wilder, S.P.; Johnson, N.; Juettemann, T.; Flicek, P.R. The ensembl Regulatory Build. *Genome Biol.* **2015**, *16*, 56. [CrossRef]
17. Hodgson, C.D.; Weeks, J.R. Symmetries, Isometries and length spectra of closed hyperbolic three-manifolds. *Exp. Math.* **1994**, *3*, 261–274. [CrossRef]
18. Gallo, F.T.; Kathe, C.; Morici, J.F.; Medina, J.H.; Weisstaub, N.V. Immediate early genes, memory and psychiatric disorders: Focus on c-Fos, Egr1 and Arc. *Front. Behav. Neurosci.* **1998**, *12*, 79. [CrossRef] [PubMed]
19. Glover, J.N.; Harrison, S.C. Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature* **1995**, *373*, 257–261. [CrossRef]
20. Hashimoto, H.; Olanrewaju, Y.; Zheng, Y.; Wilson, G.G.; Zhang, X.; Cheng, X. Wilms tumor protein recognizes 5-carboxylcytosine within a specific DNA sequence. *Genes Dev.* **2019**, *28*, 2304–2313. [CrossRef]
21. Nair, S.K.; Burley, S.K. X-ray structures of Myc-Max and Mad-Max recognizing DNA: Molecular bases of regulation by proto-oncogenic transcription factors. *Cell* **2003**, *112*, 193–205. [CrossRef]
22. Zeeman, E.C. Linking spheres. *Abh. Math. Sem. Univ. Hamburg* **1960**, *24*, 149–153. [CrossRef]
23. Rolfsen, D. *Knots and Links*; AMS Chelsea Publishing: Providence, RI, USA, 2000.
24. Schaeffer, L.N.; Huchet-Dymanus, M.; Changeux, J.P. Implication of a multisubunit Ets-related transcription factor in synaptic expression of the nicotinic acetylcholine receptor. *EMBO J.* **1998**, *17*, 3078–3090. [CrossRef]
25. Nguyen, T.T.; Grimm, S.A.; Bushel, P.R.; Li, J.; Li, Y.; Bennett, B.D.; Lavender, C.A.; Ward, J.M.; Fargo, D.C.; Anderson, C.W.; et al. Revealing a human p53 universe. *Nucl. Acids Res.* **2018**, *46*, 8153–8167. [CrossRef]
26. Nakamivhi, N.; Yoneda, Y. Transcription factors and drugs in the brain. *Jpn J. Pharmacol.* **2002**, *89*, 337–348. [CrossRef] [PubMed]
27. Chen, Y.; Zhang, X.; Dantas Machado, A.C.; Ding, Y.; Chen, Z.; Qin, P.Z.; Rohs, R.; Chen, L. Structure of p53 binding to the BAX response element reveals DNA unwinding and compression to accommodate base-pair insertion. *Nucleic Acids Res.* **2013**, *41*, 8368–8376. [CrossRef] [PubMed]

28. Baake, M.; Grimm, U. *Aperiodic Order, Volume I: A Mathematical Invitation*; Cambridge University Press: Cambridge, UK, 2013.
29. Planat, M.; Aschheim, R.; Amaral, M.M.; Fang, F.; Irwin, K. Complete quantum information in the DNA genetic code. *Symmetry* **2020**, *12*, 1993. [[CrossRef](#)]
30. 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. [[CrossRef](#)]
31. Grandy, J.K. The three neurogenetic phases of human consciousness. *J. Conscious Evol.* **2018**, *9*, 24.
32. Changeux, J.P. Allosteric receptors: From electric organ to cognition. *Annu. Rev. Pharmacol.* **2010**, *50*, 1–38. [[CrossRef](#)] [[PubMed](#)]
33. Feinberg, T.E.; Mallatt, J. The evolutionary and genetic origin of consciousness in the Cambrian Period over 500 million years ago. *Front. Psychol.* **2013**, *4*, 667. [[CrossRef](#)]
34. Amaral, M.M.; Fang, F.; Hammock, D.; Irwin, K. Geometric state sum models from quasicrystals. *Foundations* **2021**, *1*, 155–168. [[CrossRef](#)]
35. Amaral, M.M.; Fang, F.; Aschheim, R.; Irwin, K. On the Emergence of Space Time and Matter from Model Sets. *Preprint* **2021**, 2021110359. [[CrossRef](#)]