GROUP THEORY OF SYNTACTICAL FREEDOM IN DNA TRANSCRIPTION AND GENOME DECODING

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

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 structure close to free groups of rank three or less, a property that we call 'syntactical freedom'. Such a property is associated to 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 a DNA syntactical rule and their functional role. 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 role in transcription and regulation.

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

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

Keywords DNA transcription factors, genes, finitely-generated groups, free groups, aperiodic order.

1. Introduction

In his recent paper, K. I. 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 poems [3].

In the present contribution, syntactical freedom becomes the synonymous of free groups in the encoding of strings of symbols -the motifs of DNA transcription. Most transcription factors, but not all, have motifs such that 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 disfunction 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 at least 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 about 1600 binding domains in the human genome that function as transcription factors. There exist classes of DNA-binding domains of transcription factors. 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 of all introns (the non-coding regions of RNA transcript) by RNA splicing.

Promoter A sequence of DNA to which proteins initiate 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. The 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.

Free groups and their conjugacy classes. Let F_r be the free group on r generators. Following a theorem derived by Hall in 1949 [7], 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 $\mathrm{Isoc}(X;d)$ 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 [8, 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 [2],[8, Table 3.2].

Table 1. The number of conjugacy classes of subgroups of index d in the free group of rank r [8].

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 of conjugacy classes for subgroups (card seq) of a finitely generated group fp 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. But the finitely generated group $f_p = \langle x_1, x_2 | rel(x_1,x_2)\rangle$, or $f_p = \langle x_1,x_2,x_3|rel(x_1,x_2,x_3)\rangle$, or $f_p = \langle x_1,x_2,x_3,x_4|rel(x_1,x_2,x_3,x_4)\rangle$ (where the x_i are taken in the four bases A, T, G, C and rel is the motif) is not the free group $F_1 = \langle x_1,x_2|x_1x_2\rangle$, or $F_2 = \langle x_1,x_2,x_3|x_1x_2x_3\rangle$, or $F_3 = \langle x_1,x_2,x_3,x_4|x_1x_2x_3x_4\rangle$. The closeness of f_p to F_r can be checked in the finite range of indices of the card seq.

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 $H_q = \langle x_1,x_2|x_1^2 = x_2^q\rangle$ [9]. 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 $z \to -1/z$, $z \to z+\lambda_q$ 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 2, 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 polymorphism (SNP) 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, we analyze the binding domains and the card seq associated to motifs of the immediately early genes Fos, EGR1 and Myc. In such cases, the card seq of the group fp, 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. But it is important to investigate the transcription factors with a group structure away from a free group. This is done in Section 3.2 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 disfunction ot 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 conclusion, we offer some roads of progress in the connection of group theory to genetics.

2. 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. 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 [10, 11]. 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 β sheets in the protein.

The regulation of gene transcription by transcription factors depends on the gene and is governed by 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.

Gilbert's syndrome. Gilbert's syndrome is a genetic polymorphism associated to the gene UGTIA1, a phase II drug-metabolizing enzyme, which is essential in the metabolism of bilirubin and other drugs. The core promoter in UGTIA1 contains a TATA box located at position -28 [in terms of the number of amino acids (aa)] with respect to the transcriptional start site. A polymorphism with AT(TA)$_l$TAA (with l=5..8) is common in all ethnic populations with l=6 as the major allele. Minor alleles with $l > 6$ have less UGTIA1 transcription efficiency leading to Gilbert's syndrome, neonatal jaundice and toxicity in cancer chemotherapy [12].

In table 2, we looked 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 than 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 UGTIA1 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

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 UGTIA1 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 $H_q = \left\langle A, T | A^2 T^{-q} \right\rangle$ 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	$[\mathbf{1,1,2,3,2,8,7,10,18,28},\cdots]$	H_3	[6, MA0108.1]
TATAAAAA	$[\mathbf{1,3,2,8,6,19,16,69,83,238},\cdots]$	H_4	[11]
$A(TA)_5TAA$	$[\mathbf{1,3,3,7,6,34,42,123,319,706},\cdots]$	H_6	[12]
$A(TA)_6TAA$	$[\mathbf{1,1,1,1,1,1,34,77,79,51},\cdots]$	H_7	.
$A(TA)_7TAA$	$[\mathbf{1,3,2,8,6,19,16},171,315,1022\cdots]$	$\approx H_4$	.
$A(TA)_8TAA$	$[\mathbf{1,1,2,3,2,8,7,10},308,792\cdots]$	$\approx H_3$	.

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 to an imperfect fit of the group G to a Hecke group.

Single nucleotide polymorphism. The canonical form of TBP- binding site, the TATA box, is the best-studied regulatory element among human gene promoters. Tables identifying single nucleotide polymorphism (SNP) in the (gene dependent) TATA box have been collected in Reference [13].

At present, there are about 10^8 stored SNP markers that have been identified in the human genome and about 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. But regulatory SNPs in the TATA regions have biomedical usefulness and are correctable by medication and/or lifestyle. Ref. [13] 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 explicit the cardinality sequence of cc of subgroups (card seq). 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 asterisc * corresponds to the only two-base SNP marker in the table. In this case, the card seq is the same than 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 [15].

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, 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.

3. A few DNA/protein complexes an their transcription factors

3.1. **Immediate early genes and their motifs.** Immediate early genes (IEGs) are genes which 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 naturally 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 development of psychiatric disorders [14].

The group structure of the motifs of some IEG's in the Fos, EGR and Myc classes is summarized in table 4.

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 $5' - \mathrm{TGAGTCA} - 3'$ DNA elements [16]. 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 Fig. 1.

Let consider the group $fp = \langle \mathrm{A,T,G,C}|\mathrm{rel})\rangle$ on 4 letters with the relation rel = bind = TGAGTCA. The card seq of f_p up to index 6 is that of the free group $F_3 = \langle \mathrm{A,T,G,C}|\mathrm{AGTC})\rangle$ 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 f_p and F_3 are the same. This shows that both groups are close, at least in the finite range of subgroups. But fp and F_3 are not the same group. Incidentally, the group $fp' = \langle \mathrm{A,T,G,C}|\mathrm{AGTC},\mathrm{bind})\rangle$, 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 $fp = \langle \mathrm{A,T,G,C}|\mathrm{rel}\rangle$, 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:1F0S), have the same card seq than F_3 up to index 6.

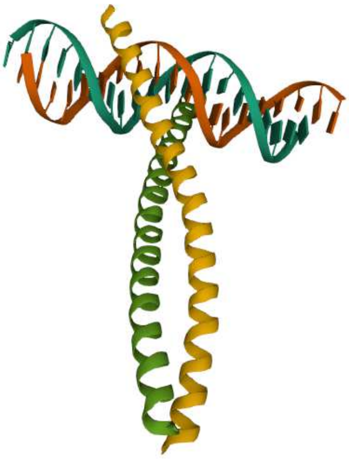


FIGURE 1. The DNA binding domain of the immediate early gene Fos. The name in the protein data bank is 1FOS.

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 $\mathrm{Cys_2His_2}$-like fold structure of a zinc finger as shown in Fig 2. It binds to the motif 5'-bind-3' [17], with bind=GCG(T/G)GGGCG.

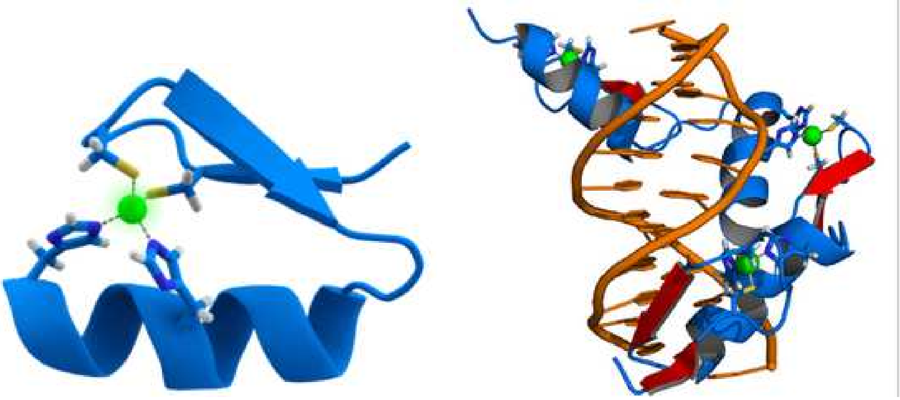


FIGURE 2. (Left) Cartoon representation of the $\mathrm{Cys_2His_2}$ zinc finger motif, consisting of an α-helix and an antiparallel β-sheet. The zinc ion (green) is coordinated by two histidine residues and two cysteine residues. (Right) Cartoon representation of the protein ZNF268 (blue) containing 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 target genes 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. A 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=TACGCCCACGC.

As for Fos domain above, let consider the group $fp = \langle \text{A,T,G,C}|\text{rel})\rangle$ on 4 letters with the relation bind or rel. The card seq of fp up to index 6 is simlar to that of the free group $F_3 = \langle \text{A,T,G,C}|\text{ATGC})\rangle$ 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 f_p and F_3 are the same. This shows that both groups are close, at least in the finite range of subgroups. But 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 $fp' = \langle \text{A,T,G,C}|\text{ATGC}, \text{bind})\rangle$, 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 its 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 ERPYACPVESCDRRFSRSDELTRHIRIHTG QKPFQCRICMRNFSRSDHLTTHIRTHTG EKPFACDICGRKFARSDERKRHTKIHLR 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 α-helix segment, letter E is for the β-sheet segment and letter C is for the random coil segment)

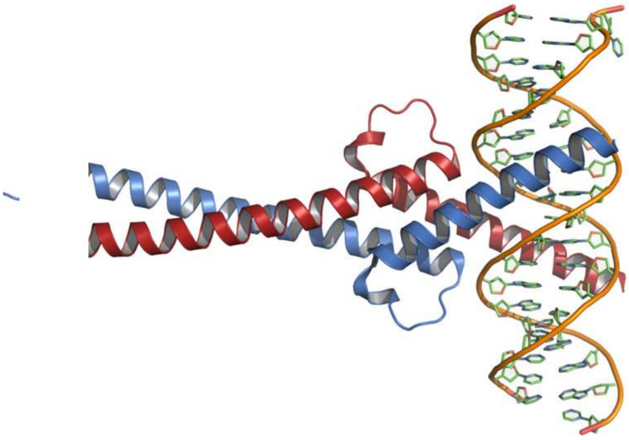
CCCEECCCCCCCCEECHHHHHHHHHHHHHH CCCEECCCCCCEECHHHHHHHHHHHHH CCCEECCCCCCEECHHHHHHHHHHHHC

Taking the former 3-letter chain as the relation of a finitely generated group on 3 letters (and rank 2), one gets the cardinality sequence for the cc of its subgroups as $[1, 3, 7, 26, 112, 717, \cdots]$ which fits the cardinality sequence of cc of subgroups of the free group F_2 only up to the index 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 [18]. 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 Fig. 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 consider the group $fp = \langle \text{A,T,G,C}|\text{rel})\rangle$ on 4 letters with the relation rel = bind = CACGTG. The conjugacy classes of fp up to index 6



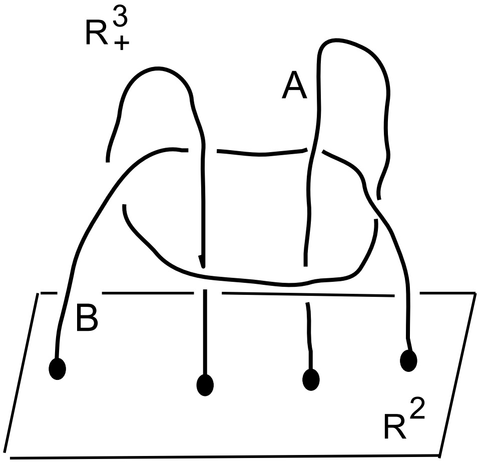


FIGURE 3. (Up) Crystal structure of Myc and Max in complex with DNA. (Down) The link $L = A \cup B$ (that is supposed to control the binding domain Myc) is attached to the plane R^2 in the half-space R^3_+ . It is not splittable. This can be proved by checking that the fundamental group $\pi = \pi_2(L)$ is not free [19],[20, 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 π_2 is $[1, 3, 10, 51, 164, 1365, 9422, 81594, 721305, \cdots]$.

have card seq equal to that of the free group $F_3 = \langle$A,T,G,C$|$GACT$)\rangle$ of rank 3. One can use again 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 f_p and F_3 are the same. This shows that both groups are close, at least in the finite range of subgroups. But fp and F_3 are not the same group. The group $fp' = \langle \text{A,T,G,C}|\text{GACT}, \text{bind})\rangle$, with the joint relations of fp and F_3, is close to the group π_2 defined in Fig. 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 that are not splittable. The proof is in Refs [19],[20, p. 90] and follows from the fact that π_2 is not a free group. The group fp' is close to π_2 in the sense that the cardinality sequence $[1, 3, 10, 51, 164, 1365, 9422, 81594, 721305, \cdots]$ of the cc of subgroups is that of π_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 π_2 is not free are distinguished features of the Myc domain that it is tempting to associate to a potential abnormal replication.

3.2. 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 are not leading to free groups. This is shown in Table 5. A short account of the function or disfunction of the corresponding genes is in Table 6. It is observed that several transcription factors whose group structure is a 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].

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 p53 plays in preventing cancer development. The p53 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 [22], 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 π'_3. But there are motifs leading to a card seq associated to the free group F_3 or to other non-free groups that are not of type π'_3. That may be due to the fact that p53 has many isoforms to fill its role.

In Figure 4, we borrow the crystal structure of the p53 domain for the binding domain of the PDB sequence 4HJE. The p53 domain forms a tetramer. But other symmetries of the binding domain of p53 may be found.

4. Syntactical freedom and aperiodicity

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?

We first introduce the concept of a general substitution rule in the context of free groups. A general substitution rule ρ on a finite alphabet $\mathcal{A}_r$ on r letters is an endomorphism of the corresponding free group F_r [24, 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 $u, v \in F_r$.

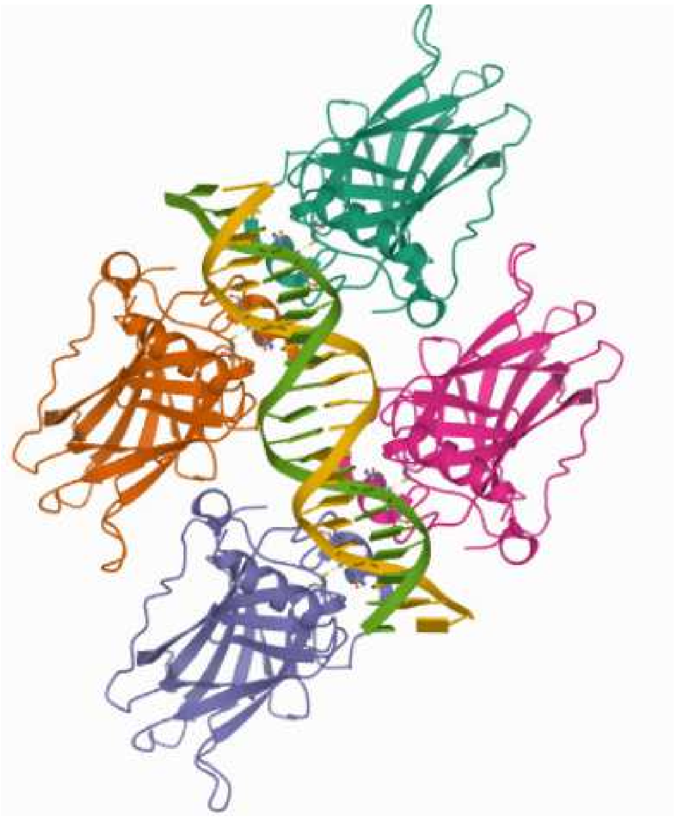


FIGURE 4. Crystal structure of p53 binding domain. The reference number in the protein data bank is 4HJE.

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

The map α induces a homomorphism $M : \mathrm{End}(F_r) \to \mathrm{Mat}(r, \mathbb{Z})$. Under M, $\mathrm{Aut}(F_r)$ maps to the general linear group of matrices with integer entries $GL(r, \mathbb{Z})$. Given ρ, there is a unique mapping $M(\rho)$ that makes the map diagram commutative [24, p. 68]. The substitution matrix $M(\rho)$ of ρ may be specified by its elements at row i and column j as follows

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

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

Then, we are interested in the sequence of finitely generated groups

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

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.

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 $M \in \mathrm{Mat}(d, \mathbb{R})$ 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 $M >> 0$) has all its entries positive. An irreducible matrix $M = (M_{ij})_{1 \le i,j \le d}$ 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 [24, Corollary 4.3]. A primitive substitution rule ρ of substitution matrix $M(\rho)$ with an irrational PF-eigenvalue is aperiodic.

A well studied primitive substitution rule is The Fibonacci rule $\rho = \rho_F : \; a \to ab, \; b \to a$ of substitution matrix $M_F = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ and PF-eigenvalue equal to the Golden ratio $\lambda_{PF} = \tau = (\sqrt{5} + 1)/2$ [24, Example 4.6]. As expected, the irrationality of λ corresponds to the aperiodicity of the Fibonacci sequence.

The sequence of Fibonacci words is as follows

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

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

It is straightforward to check that all finitely generated groups $f_p^{(l)}$ whose relations $\mathrm{rel}(a, b) = ab, aba, abaab, abaababa, \cdots$ 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.

Let us now provide examples taken for transcription factors involving 2, 3 or 4 letters.

A two-letter sequence for the transcription factor of gene DBX in drosophila melanogaster. Let us consider the motif rel=TTTATTA for the gene DBX in drosophila melanogaster (fruit fly) [6, MA0174.1]. The roles of the DBX gene include neuronal specification and differentiation.

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

doi:10.20944/preprints202112.0381.v2

$$A, T, AT, TTTATTA, TTATTATTATTTATTATTATTTA \cdots$$

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

Similarly to the Fibonacci generator rules, all finitely generated groups $f_p^{(l)}$ whose relations are $\mathrm{rel}(A, T) = AT, TTTATTA, TTATTATTATTTATTATTATTTA \cdots\cdots$ 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.

A three-letter sequence for the transcription factor of gene EGR1. The transcription factor of gene EGR1 was investigated in Section 3.1. The selected motif is rel=GCGTGGGCG. We split rel into two segments so that $\mathrm{rel} = \mathrm{rel}_C \mathrm{rel}_G \mathrm{rel}_T$ with the substitution maps $C \to \mathrm{rel}_C = G$, $G \to \mathrm{rel}_G = CGT$, $T \to \mathrm{rel}_T = GGGCG$ to produce the substitution sequence

$$C, G, T, CGT, GCGTGGGCG, CGTGCGTGGGCGCGTCGTCGTGCGT, \cdots$$

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 $M^2 >> 0$) whose eigenvalues follow from the vanishing of the polynomial $-\lambda^3 + \lambda^2 + 5\lambda + 1 = 0$. There are three real irrational roots $\lambda_1 \approx 2.86619$, $\lambda_2 \approx -0.21075$ a,d $\lambda_3 \approx -1.65544$. 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 $\mathrm{rel}(C, G, T) = CGT, GCGTGGGCG, CGTGCGTGGGCGCGTCGTCGTGCGT, \cdots$ have a card seq whose elements are $1, 3, 7, 26, 97, 624, 4163 \cdots$ which is the card seq of the free group F_2.

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

A four-letter sequence for the transcription factor of the Fos gene. The transcription factor of gene Fos was investigated in Section 3.1. The selected motif is rel=TGAGTCA.

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

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

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 primitive matrix ($M^4 >> 0$) 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 $\mathrm{rel}(A, C, G, T) = ATGC, TGAGTCA, GAGTCTAGTCGAT, \cdots$ have a card seq whose elements are $1, 7, 41, 604, 13753, 504243 \cdots$. 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. Conclusion

We made use of group theory for investigating transcription factors in genetics. Finite group theory plays a big role in the attempts to model the genetic code, see [25, 26] 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 noticed that many structures for the protein secondary codes tend to be close to free groups. Of course, 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 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 less than 10 amino acids. Taking random sequences instead of the gene specific DNA sequences in TFs also lead 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. But motifs in TFs are codes with a particular meaning - the specific gene function or disfunction. In this sense, we found appropriate to use the concept of 'syntactical freedom' to qualify most TFs and to associate the lack of syntactical freedom to potential source of illness. In our context, syntactical freedom means free groups and aperiodicity.

An interesting line of research is about the neurogenetic correlates of human consciousness and the related TFs. There are some papers in this direction [21], [27]-[29].

As a final note, we refer to the paper [30, 31] 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.

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.

[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.

[3] Planat, M.; Aschheim, R.; Amaral, M. M.; Fang F.; Irwin K. Graph coverings for investigating non local structures in protein, misic and poems, *Sci* **2021** 3, 39.

[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. The classication is available at http://www.edgar-wingender.de/huTF_classification.html.

[5] Wingender, E.; Schoeps, T.; Dönitz, J. TFClass: An expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* **2013** T1, D165–D170.

[6] Sandelin, A; Alkema, W; Engström, P; Wasserman, WW; Lenhard, B, JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research* **2004** 32, pp D91–D94; software available at https://jaspar.genereg.net/

[7] M. Hall Jr, Subgroups of finite index in free groups, *Can. J. Math.* **1949** 1, 187–190.

[8] Kwak, J. H.; Nedela R. Graphs and their coverings, *Lecture Notes Series* **2007** 17, 118 pp.

[9] The modular group, available at https://en.wikipedia.org/wiki/Modular_group, accessed on 1 october 2021.

[10] TATA box, https://en.wikipedia.org/wiki/TATA_box, accessed on 1 September 2021.

[11] 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 Research* **1996** 24, 3100–3106.

[12] 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.

[13] 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.; Kolchanov, N. A.; Osadchuk, L. V.; Osadchuk, A. V. Candidate SNP markers of reproductive potential are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *BMC Genomics* **2018** 19, 16–38.

[14] Gallo, F. T.; Katche C.; Morici J. F.; Medina J. H.; Weisstaub N. V. Immediate early genes, memory and psychiatric disorders: focus on c-Fos, Egr1 and Arc. *Frontiers in behavioral neuroscience* **1998** 12, 79.

[15] Hodgson, C. D.; Weeks J. R. Symmetries, Isometries and length spectra of closed hyperbolic three-manifolds. *Exp. Math.* **1994** 3, 261–274.

[16] 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.

[17] Hashimoto, H.; Olanrewaju Y. 0.; 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.

[18] 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.

[19] Zeeman E. C. Linking spheres. *Abh. Math. Sem. Univ. Hamburg* **1960** 24, 149–153.

[20] Rolfsen D. *Knots and Links*; AMS Chelsea Publishing: Providence, Rhode Island, USA, 2000.

[21] 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.

[22] 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.; Li, L.; Resnick, M. A.; Menendez D. Revealing a human p53 universe. *Nucl. Ac. Res.* **2018** 46, 8153–8167.

[23] Nakamivhi, N.; Yoneda Y. Transcription factors and drugs in the brain. *Jpn J. Pharmacol* **2002** 89, 337–348.

[24] Baake M., Grimm U. *Aperiodic order, Vol. I: A mathematical Invitation*; Cambrige Univ. Press: Cambridge, UK, 2013.

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

[26] 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.

[27] Grandy, J. K. The three neurogenetic phases of human consciousness. *J. Conscious Evolution* **2018** 9, 24pp.

[28] Changeux J. P. Allosteric receptors: from electric organ to cognition. *Annu. Rev. Pharmacol* **2010** 50, 1–38.

[29] in the Cambrian Period over 500 million years ago. *Frontiers in Psychology* **2013** 4, 667. Feinberg T. E.; Mallatt, J. The evolutionary and genetic origin of consciousness

[30] Amaral, M.M.; Fang, F.; Hammock D.; Irwin K. Geometric state sum models from quasicrystals. *Foundations* **2021** 1, 155–168.

[31] Amaral, M.M.; Fang, F.; Aschheim R.; Irwin K. On the emergence of space time and matter from model sets. Prepint 2021, 2021110359 (doi: 10.20944/preprints202111.0359.v1).

doi:10.20944/preprints202112.0381.v2

TABLE 3. Group analysis of a few known and candidate SNP markers (taken from [13]). Three-base markers are taken into account. 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 worth (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 [13]. The asterix * corresponds to the only two-base SNP marker in the table. The card seq is the same than the sequence for the fundamental group of 3-manifold $m002$. The latter manifold is the smallest volume closed 3-manifold and is non orientable [15].

gene	rel: marker	card seq of cc of subgroups	Litterature
ESR2	TTAAAAGGAA	$[\mathbf{1}, 7, 17, 114, 423, 4526, 30364, 293306 \cdots]$	[13, Tab. 1], B
HSD17B1	AGCCCAGAGC	$[\mathbf{1}, \mathbf{3}, \mathbf{7}, \mathbf{26}, 217, 124, 18443, 219870 \cdots]$	., A
.	CAAGCCCAGA	$[\mathbf{1}, 7, 14, 109, 396, 3347, 19758, 188940 \cdots]$	., A
PGR	AAAGGAGCCG	$[\mathbf{1}, 7, 17, 142, 475, 4125, 23509, 225871 \cdots]$	., A
GSTM3	GGGTATAAAG	$[\mathbf{1}, 7, 14, 109, 396, 3347, 19758, 188940 \cdots]$	., E
.	CCCCTCCCGC	$[\mathbf{1}, \mathbf{3}, \mathbf{7}, \mathbf{26}, \mathbf{97}, \mathbf{624}, \mathbf{4163}, \mathbf{34470} \cdots]$	., C
.	CCCTCCCGCT	.	., C
IL1B	AAAACAGCGA	$[\mathbf{1}, 7, 14, 89, 224, 1842, 10191, 86701 \cdots]$	[13, Tab. 2], A
CYP2A6	AAAGGCAAC	$[\mathbf{1}, 7, 17, 134, 683, 7077, 64225 \cdots]$	., A
DHFR	GGGACGAGGG	$[\mathbf{1}, \mathbf{3}, \mathbf{7}, \mathbf{26}, \mathbf{97}, \mathbf{624}, \mathbf{4163}, \mathbf{34470} \cdots]$	., A
.	GGACGAGGGG	.	., A
LEP	GGGGCGGGA	$[\mathbf{1}, \mathbf{3}, \mathbf{7}, \mathbf{26}, \mathbf{97}, \mathbf{624}, \mathbf{4163}, \mathbf{34470} \cdots]$	[13, Tab. 3], C
GCG	TGCGCCTTGG	$[\mathbf{1}, \mathbf{3}, \mathbf{7}, \mathbf{26}, 119, 816, 4865, 40489 \cdots]$	., B
GH1	TATAAAAAGG	$[\mathbf{1}, 7, 14, 109, 396, 3347, 19758, 188940]$	., E
.	GTATAAAAAG	.	., D
.	GGTATAAAAA	.	., E
.	AGGGCCCACA	$[\mathbf{1}, \mathbf{3}, \mathbf{7}, \mathbf{26}, 127, 860, 5661, 45710 \cdots]$	., A
.	AAAGGGCCCC	$[\mathbf{1}, \mathbf{3}, 10, 67, 266, 3458, 30653, 312237 \cdots]$	., A
.	AAAGGGCCA	.	., A
NOS2	TCTTGGCTGC	$[\mathbf{1}, \mathbf{3}, \mathbf{7}, \mathbf{26}, \mathbf{97}, \mathbf{624}, \mathbf{4163}, \mathbf{34470} \cdots]$	[13, Tab. 4], A
TPI1	ATATAAGTGG	$[\mathbf{1}, \mathbf{3}, \mathbf{7},30, 125, 856, 4832, 40246 \cdots]$	., B
GJA5	TATTAAACAC	$[\mathbf{1}, \mathbf{3},10, 35, 140, 921, 5778, 47238 \cdots]$	., E
HBD	AAAAGGCAGG	$[\mathbf{1}, \mathbf{3}, \mathbf{7}, \mathbf{26}, \mathbf{97}, \mathbf{624}, \mathbf{4163}, \mathbf{34470} \cdots]$	[13, Tab. 5], A
F2	AACCCAGAGG	$[\mathbf{1}, \mathbf{3}, \mathbf{7}, \mathbf{26},127, 860, 5661, 45710 \cdots]$	., A
F8	GGAAGAGGGA	$[\mathbf{1}, \mathbf{3},2, 7, 4, 18, 9, 27, 36, 68 \cdots]^*$	[13, Tab. 6], A
F3	GCGCGGGGCA	$[\mathbf{1}, \mathbf{3}, \mathbf{7}, \mathbf{26}, \mathbf{97}, \mathbf{624}, \mathbf{4163}, \mathbf{34470} \cdots]$	., A
F11	TTTTTAGTAA	.	., D
.	TTTTTAGTAA	$[\mathbf{1},7, 17, 114, 423, 4526, 30364, 293306 \cdots]$	., A
.	AAGGAAATTT	$[\mathbf{1}, \mathbf{3}, \mathbf{7}, \mathbf{26},195, 1692, 11803, 73192 \cdots]$	., A
AR	GTGGAAGATT	$[\mathbf{1}, \mathbf{3}, \mathbf{7},34, 139, 931, 5208, 43867 \cdots]$	[13, Tab. 7], A
.	CCACGACCCG	$[\mathbf{1},7, 20, 167, 754, 7232, 60860, 683597 \cdots]$	., D
MTHFR	TCCCTCCCA	$[\mathbf{1}, \mathbf{3}, \mathbf{7}, \mathbf{26}, \mathbf{97}, \mathbf{624}, \mathbf{4163}, \mathbf{34470} \cdots]$	., A
DMNT1	TGTGTGGCCCG	.	., A
.	GTGTGTGCCC	.	., A
.	GACGAGCCCA	$[\mathbf{1}, \mathbf{3}, \mathbf{7},42, 131, 912, 6011, 47322 \cdots]$	., A
NR5A1	ACAAGAGAAA	$[\mathbf{1}, \mathbf{3}, \mathbf{7}, \mathbf{26}, \mathbf{97}, \mathbf{624}, \mathbf{4163}, \mathbf{34470} \cdots]$	., A
.	GGTGTGAGAG	$[\mathbf{1},7, 14, 89, 264, 1987, 11086, 93086 \cdots]$	., A

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 $[1,3,2,5,2,7,2,8,3,8,2,13,2,9,4,\cdots]$.

gene	rel: motif	card seq	Litterature
Fos	TGAGTCA	F_3	[16]
.	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	CCGCCCACGC	.	., MA0472.1
.	ACGCCCACGCA	.	., MA0472.2
EGR3,EGR4	ACGCCCACGCA	.	., [MA0732.1, MA0733.1]
Myc	CACGTG	F_3	[16]
.	CGCACGTGGT	.	[6], MA0147.1
.	CCCACGTGCTT	.	., MA0147.2
.	CCACGTGC	.	., MA0147.3
Mycn, Max::Myc, etc	GACCACGTGGT, etc	.	., [MA0104.1, etc]

TABLE 5. Group structure of motifs for some transcription factors that are not leading to free groups. The card seq for π_1 is $[1,4,1,2,4,2,1,7,2,2,4,2,2,8,1,2,7,2,3,\cdots]$; for π_1' it is $[1,1,1,2,1,3,3,1,2,2,1,1,9,2,14,2,1,\cdots]$. The card seq for π_2 is already in Fig. 3 as $[1,3,10,51,164,1365,9422,81594,721305,\cdots]$. The card seq for π_3 is $[1,7,14,89,264,1987,11086,93086,\cdots]$; for π_3' it is $[1,7,50,867,15906,570528,\cdots]$; for π_3'' it is $[1,7,50,739,15234,548439,\cdots]$; for $\pi_3^{(3)}$ it is $[1,7,41,668,14969,550675]$. The card seq for π_4 is $[15,82,1583,30242\cdots]$. The index i in π_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	Litterature
NKX6-2	TAATTAA	H_3	[6], [MA0675.1, MA0675.2]
HoxA1, HoxA2	TAATTA	π_1	[6], [MA1495.1, MA0900.1]
POU6F1, Vax	.	.	., [MAO628.1, MA0722.1]
RUNX1	TGTGGT	.	., MA0511.1
RUNX1	TGTGGTT	π_1'	[6], MA0002.2
EHF	CCTTCCTC	.	., MA0598.1
POU6F1	TAATGAG	π_2	[6] MA1549.1
PITX2	TAATCCC	.	., [MA1547.1, MA1547.2]
ELK4	CTTCCGG	.	., MA0076.2
OTX2, Dmbx1	GGATTA	π_3	[6], [MA0712.2, MA0883.1]
PitX1, PitX2, PitX3, OTX1	TAATCC	.	.,[MA0682.1, MA0711.1]
N-box	TTCCGG	.	[21]
p53	CACATGTCCA	π_3'	[22]
GZF1	TGCGCGTCTATA	.	[4]
NF-kappa-B	GGGAATTTCC	.	[6], [MA0107.1, MA1911.1]
STAT1	TTTCCCGGAA	.	., MA0137.2
.	TTCCAGGAA	.	., MA0137.3
STAT4	TTCCAGGAAA	.	., MA0518.1
FOSL1::Jun	ATGACGTCAT	π_3''	[6], MA1129.1
USF2	GTCATGTGACC	.	. , MA0626.1
PAX1	CGTCACGCATGA	.	. , MA0779.1
STAT2	TTCCAGGAAG	.	. , MA0144.1
FOS	GATGACGTCATCA	$\pi_3^{(3)}$	[6], MA1951.1
MAFA, MAFF,MAFK	TGCTGAGTCAGCA	.	., [MA1521.1, MA0495.2, MA0946.2]
CREB	TGACGTCA	π_4	[6], [MA0018.2, MA018.3]
USF2	GGTCACGTGACC	.	., MA0526.4
SMAD3, SMAD5	GTCTAGAC	.	., [MA0795.1, MA1557.1], [23]

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

gene	type	function	disfunction
NKX6-2	homeobox	central nervous system, pancreas	spastic ataxia
HoxA1	homeobox	embryonic devt of face and hear	autism
HoxA2	.	.	cleft palate
Pou6F1	.	neuroendocrine system	clear cell adenocarcinoma
Vax	.	forebrain development	craniofacial malform.
RunX1	Runt-related	cell differentiation, pain neurons	myeloid leukemia
EHF	homeobox	epithelial expression	carcinogenesis, asthma
PitX2	.	eye, tooth, abdominal organs	Axenfeld-Rieger syndrome
ELK4	Ets-related	serum response for c-Fos	
OTX1,OTX2	homeobox	brain and sensory organ devt	medulloblastomas
Dmbx1	.	.	farsightedness and strabismus
PitX1	.	organ devt, left-right asymmetry	autism, club foot
PitX3	.	lens formation in eye	congenital cataracts
N-box	Ets-related	synaptic expression	drug sensitivity
p53	p53 domain	‘Guardian of the genome’	cancers
GZF1	Zinc fingers	protein coding	short stature, myopia
NF-kappa-B	.	DNA transcription, cytokines	apoptosis
STAT1	Stat family	signal activator of transcription	immunodeficiency 31
STAT4	Stat family	signal activator of transcription	rheumatoid arthritis
FOSL1::Jun	leucine zipper	cellular proliferation	marker of cancer
USF2	helix-loop-helix	transcription activator	
PAX1	paired box	fetal development	Klippel–Feil syndrome
FOS	leucine zipper	cellular proliferation	cancers
Maf	.	pancreatic development	congenital cerulean cataract
CREB	bZIP	neuronal plasticity and long term memory	Alzheimer’s disease
USF2	helix-loop-helix	transcription activator	
SMAD	homeo domain	regulation of cell development and growth	Alzheimer’s disease