Journal of Theoretical Biology 523 (2021) 110708

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Motif lengths of circular codes in coding sequences

M. Gumbel *, P. Wiedemann

Competence Center for Mathematical and Algorithmical Methods in Biology, Biotechnology and Medicine, Mannheim University of Applied Sciences, 68163 Mannheim, Germany

ARTICLE INFO

Article history:
Received 15 May 2020
Revised 22 March 2021
Accepted 30 March 2021
Available online 20 April 2021

Keywords:
Coding region
Circular codes
Motif

ABSTRACT

Protein synthesis is a crucial process in any cell. Translation, in which mRNA is translated into proteins, can lead to several errors, notably frame shifts where the ribosome accidentally skips or re-reads one or more nucleotides. So-called circular codes are capable of discovering frame shifts and their codons can be found disproportionately often in coding sequences. Here, we analyzed motifs of circular codes, i.e. sequences only containing codons of circular codes, in biological and artificial sequences. The lengths of these motifs were compared to a statistical model in order to elucidate if coding sequences contain significantly longer motifs than non-coding sequences. Our findings show that coding sequences indeed show on average greater motif lengths than expected by chance. On the other hand, the motifs are too short for a possible frame shift recognition to take place within an entire coding sequence. This suggests that as much as circular codes might have been used in ancient life forms in order to prevent frame shift errors, it remains to be seen whether they are still functional in current organisms.

1. Introduction

The universal genetic code defines the way genetic information stored in DNA is translated into proteins. It comprises a total of 64 codons; the three positions in a codon can each be occupied by one of the four bases G, A, T and C. Overall, many factors contribute to the fidelity of this translation process. The error rate in replicating genomic DNA is as low as 10^{-9} to 10^{-11} errors per base pair (in *E. coli*; Fijalkowska et al., 2012). Errors in transcription have been estimated to occur at a rate of 10^{-4} to 10^{-5} per nucleotide in bacteria Meyerovich et al., 2010 and 10^{-5} to 10^{-6} per nucleotide in eukaryotic cells (Drummond and Wilke, 2009). More error prone is the actual translation of codons into a growing amino acid chain in the ribosome. Here, error rates of 10^{-3} to 10^{-4} per codon have been reported in bacteria (Meyerovich et al., 2010).

During translation, the ribosome must ensure that the mRNA strand is pushed forward for exactly three nucleotides at a time in order to keep the reading frame as determined by the start codon. If something goes wrong here, there is a frame shift resulting in very different amino acids. How can the ribosome process the RNA strand with high accuracy? One possible explanation is selection of specific tRNA repertoires that may operate to reduce frame-shifting errors (Warnecke et al., 2010). Other mechanisms have been described, e.g. avoiding frame shifts during initial stages of translation in bacteria by EF-P and m1G37 methylation of tRNA (Gamper et al., 2015). Generally, translation errors like frame shifts are likely to develop from non-canonical interactions between the mRNA codon and the tRNA anticodon, whose cognate interactions are required for accurate tRNA selection and movement of tRNAs through the ribosome during elongation (Dunkle and Dunham, 2015). One of the oldest theories was proposed by Crick et al. (1957). It claims that only 20 out of 64 codons are used in coding RNA sequences and that these 20 codons are unique in each reading frame. Each of these 20 codons would code for one of the 20 amino acids. If a frame shift occurs, the ribosome would immediately encounter *unknown* codons and terminate the translation. Such a code is called a *comma-free code*. Although this idea seems to be reasonable, there is no experimental support for those comma-free codes.

More recently another code was discovered which is supported by experimental data. Arquès and Michel analyzed 13,686 coding regions in prokaryotes and 26,757 in eukaryotes (Arquès and Michel, 1996, repeated with more sequences in Michel (2015)) and counted the frequencies of all 64 codons in these regions in frame 0, 1 and 2. Frame 0 refers to the reading frame. It turned out that the frequencies were different in each frame and that each codon had a clear maximal frequency in exactly one frame. The codons

$$X_0 = \{\mathrm{AAC}, \mathrm{AAT}, \mathrm{ACC}, \mathrm{ATC}, \mathrm{ATT}, \mathrm{CAG}, \mathrm{CTC}, \mathrm{CTG}, \mathrm{GAA}, \mathrm{GAC}, \mathrm{GAG}, \mathrm{GAT}, \mathrm{GCC}, \mathrm{GGC}, \mathrm{GGT}, \mathrm{GTA}, \mathrm{GTC}, \mathrm{GTT}, \mathrm{TAC}, \mathrm{TTC}\}$$

* Corresponding author.
E-mail addresses: m.gumbel@hs-mannheim.de (M. Gumbel), p.wiedemann@hs-mannheim.de (P. Wiedemann).

https://doi.org/10.1016/j.jtbi.2021.110708
0022-5193/

are preferably used in frame 0 and called X_0 code (0 like frame 0). Moreover, this code also contains its self-complementary and reversed codon for each codon, i.e. as AAC is in X_0 so there is GTT. It was shown by El Soufi and Michel that the coding regions of eukaryotes indeed contain more codons from X_0 than any other regions in the genome (El Soufi and Michel, 2016). A recent paper that analyzed motifs of the X_0 code in *S. cerevisiae* and other species suggests that this X_0 code may be an evolutionary relic of a primitive code originally used for gene translation (Gopal Dila et al., 2019).

Surprisingly, X_0 has the properties of a so-called *circular code* (Fimmel et al., 2015). Like comma-free codes, circular codes are capable of detecting a frame shift error. A set of (different) codons X is called a (trinucleotide) circular code if every sequence of bases written on a circle, i.e. the next letter after the last letter being the first letter, has at most one decomposition into codons from X.

The major difference to comma-free codes is that a frame shift cannot be detected immediately – depending on the circular code a frame shift is recognized not later than after reading one to four codons in the wrong frame (see also Michel, 2012). Fig. 1 gives an example.

Although codons of the circular code X_0 are over-represented in coding sequences and X_0 is capable of detecting a frame shift, the concept of circular codes also imposes strong restrictions. First of all, there exist – in theory – many circular codes and any circular code can consist of maximal 20 codons (Fimmel and Strüngmann, 2018). Like comma-free codes, they must not contain the codons AAA, UUU, CCC and GGG and they cannot contain tuple-wise shifted codons – otherwise a frame shift could never be detected. For example, if AAC is part of the code, then ACA and CAA must be omitted. Furthermore, the fact that any circular code can contain maximally 20 codons is problematic since coding sequences typically contain all 64 codons.

In order to detect any frame shift a sequence should consist only of codons of a circular code. A consecutive sequence of codons from a circular code is called a *motif* (see Table 3). Previous work has primarily analyzed the *maximal* motif lengths which can be found in coding regions (e.g. El Soufi and Michel, 2016; Gopal Dila et al., 2019). Average or even minimal motifs were not in focus. Evidently, it is likely (and we will show) that a coding sequence is interrupted many times by codons not being part of a circular code. Thus, a frame shift cannot always be detected in coding sequences – simply because the sequence does not fully consist of codons of a circular code. However, the latter might have been the case in very old sequences of ancient organisms – maybe in species as old as the genetic code itself. Mutations might have perforated ancient sequences with codons not being part of circular codes. If the circular code's ability to detect frame shifts was (or still is) applied in the translation process of sequences into proteins, one would expect that the lengths of motifs in nucleotide sequences of present-day organisms are longer (on average) than would be expected at random. This paper searches for those traces of motif lengths in coding sequences of present-day organisms.

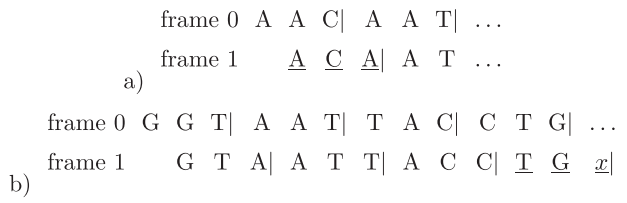

Fig. 1. Two examples for frame shift detection with code X_0. The top sequence of each example consists of codons from X_0 in frame 0, the bottom sequence is the same sequence in frame 1. a) The frame shift is recognized immediately as the codon ACA in frame 1 is not part of X_0. b) The frame shift is detected after reading the fourth codon in the wrong frame. GTA, ATT and ACC are still part of X_0, so the frame shift is not detected immediately. The next codon, however, does not belong to X_0 as there is no codon starting with TG – the frame shift is recognized. x represents any base.

The software used in this paper was written in R and the package *ccmotif* is available for download at https://github.com/information-mannheim/ccmotif.

2. Methods

Our analysis consists of the following steps: the coding sequences for all model organisms (as defined in Section 2.1) are analyzed. We take a set ($\mathcal{C}$) of selected circular codes (see 2.2) and calculate the motif length for every model organism and every code. This is done for every frame. The same is performed with random codes ($\mathcal{R}$) and non-coding sequences. The motif lengths are then compared to a statistical model.

2.1. Sequence compilations

We chose relevant eu- and prokaryotic as well as archaeal model organisms to cover a broad range of biological diversity. The coding sequences of the chosen organisms were downloaded from the resources given below. Details can be found in the online supplementary material. Those sequences denote in their FASTA files Uracil (U) as Thymine (T). Our analysis always uses T, no matter if we process DNA or RNA.

- Human (*H. sapiens*)
  mRNA sequences for *Homo sapiens* were taken from the Consensus CDS (CCDS) project[1].
- Nematode (*C. elegans*)
  mRNA sequences for *Caenorhabditis elegans* were downloaded from https://www.ebi.ac.uk/ena.
- Yeast (*S. cerevisiae*)
  mRNA sequences for *Saccharomyces cerevisiae* were downloaded from https://www.yeastgenome.org.
- Plant (*A. thaliana*)
  mRNA sequences for *Arabidopsis thaliana* were downloaded from http://plants.ensembl.org/Arabidopsis_thaliana/Info/Index. Version TAIR10 is used.
- Green alga (*C. reinhardtii*)
  mRNA sequences for *Chlamydomonas reinhardtii* were downloaded from https://www.ebi.ac.uk/ena.
- Bacteria (*E. coli*)
  mRNA sequences for *Escheria coli* were downloaded from http://bacteria.ensembl.org/Escherichia_coli.
- Archaea (*P. occultum*)
  mRNA sequences for *Pyrodictium occultum* were downloaded from https://www.ebi.ac.uk/ena/browser/view/Taxon:2309.
- Archaea (*T. tenax*)
  mRNA sequences for *Thermoproteus tenax* were downloaded from https://www.ebi.ac.uk/ena/browser/view/Taxon:768679

A summary of the coding sequences is given in Table 1. As the analysis is computationally intensive, $N = 5000$ coding sequences were randomly drawn without repetitions from the complete list of CDS if an organism had more than 5000 CDS. If not the whole CDS set was taken.

Additionally, we compare motif lengths in coding sequences with a) non-coding regions and b) random sequences in order to get a better understanding of the motif lengths distribution. For simplicity, 100,000 bases from an arbitrary region on human

[1] ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/

Table 1
Overview of coding sequences. The column *CDS* (coding sequences) counts the number of CDS and *Average codons per CDS* shows the average number of codons per CDS.

Species	CDS	Average codons per CDS
H. sapiens	32554	569.3
C. elegans	33111	472.4
S. cerevisiae	5917	501.4
A. thaliana	48321	432.1
C. reinhardtii	36626	601.8
E. coli	5494	288.9
P. occultum	1612	286.7
T. tenax	2049	271.3

chromosome 1 were extracted. *H. sapiens* is just an example, in the supplementary material we also created sequences for other organisms. Please note that properties of the DNA – like GC content – in different species is not crucial as the X_0 code is the best average code for all organisms. Our analysis is about code usage and not codon usage (see Eq. (1)). Such a sequence might contain genes but this is not intended and it does not matter if so. An artificial representative RNA sequence, however, is difficult to obtain. It is known that the likelihood for the next nucleotide in a sequence depends on the current nucleotide. Table 2 shows those transition probabilities which were estimated in human chromosome 1 (Fimmel et al., April 2019). Based on these numbers we have created an artificial DNA sequence of 100,000 nucleotides in a Markovian-like way.

2.2. Codes

2.2.1. Circular codes

As mentioned before, the circular code X_0 plays an important role as it was identified to be the most common code in all species. X_0 is a maximal circular code which means it contains 20 codons. On the other hand, those 20 codons of X_0 encode only 13 different amino acids. Any circular code can code at most for 18 amino acids (Michel, 2014). There are 12,964,440 maximal circular codes and only 10 of them have codons for 18 different amino acids (Arquès and Michel, 1996). Among all maximal circular codes there are 216 codes which are also self-complementary (Michel et al., 2008). This list is denoted as $\mathcal{M}$ and the order of the list is alphabetical according to the codons; X_0 belongs to this set and is in position 23 – it gets the label $C23$. As X_0 is the best *average* code in coding sequences, we have extended our analysis to all structurally *similar* codes. Within this paper, we have tested the coding sequences with the following codes:

- The set $\mathcal{C}_0$ which is a subset of $\mathcal{M}$ such that a code does not have stop codons. Codes with a stop codon would not have any biological meaning. There are 70 of those codes ($|\mathcal{C}_0| = 70$).
- The tuple-wise shifted version of $\mathcal{C}_0 : \mathcal{C}_1$ consists of codons with one shift and $\mathcal{C}_2$ with two shifts ($b_1, b_2, b_3 \to b_2, b_3, b_1$ and $b_1, b_2, b_3 \to b_3, b_1, b_2$). It is known that these shifted codes are very frequent in frames 1 and 2 (Michel, 2015).

Table 2
Transition probabilities as measured in human chromosome 1. x is a base and y the next base in sequence. For instance, the chance that a G follows a C is rather low (4.9%).

		y			
	$P(x \to y)$	A	T	C	G
x	A	0.327	0.255	0.173	0.245
	T	0.216	0.328	0.206	0.250
	C	0.349	0.342	0.259	0.049
	G	0.288	0.242	0.211	0.260

All codes together are denoted as $\mathcal{C} = \mathcal{C}_0 \cup \mathcal{C}_1 \cup \mathcal{C}_2$. In total there are 210 codes. As we have cyclic-shifted codons, there are $3 \cdot 20 = 60$ (different) codons out of 64 used in $\mathcal{C}$ – only AAA, TTT, CCC and GGG are missing.

2.2.2. Random codes

We would like to compare these circular codes with random codes. A random code in this context is a set of 20 different codons randomly drawn from the set of all 64 codons. In particular, a random code might contain AAA, UUU, CCC, GGG and cyclic variants of a codon. Similar to the circular codes a set $\mathcal{R}$ of 210 random codes is generated.

2.2.3. Code usage

The code usage u is defined as the relative frequency of codons that belong to a code X in relation to all 64 codons.

2.3. Motif lengths

The length of a motif is defined by the maximal number of consecutive codons from the same code (see Table 3). Note that a motif in Gopal Dila et al. (2019) was defined differently: here motifs which consist of less than four codons or contain less than four different bases were omitted. This definition cannot be used in our approach as we would get many regions on a CDS which are *blank*. As we want to quantify the frame shift recognition capabilities by means of circular codes in CDS, every motif – no matter how long or what codons it consists of – has to be considered.

Let $m_{ij}^{(X,s)}$ be a motif length for motifs of a code X ($X \in \mathcal{C}$) in species s where i indicates the coding sequence ($1iN; N = 5000$) and j the motif within the coding sequence ($1jn_i; n_i$ is the number of motifs in the i-th CDS). For simplicity we write m_{ij} if the code and organism is known. Note that n_i is in general not a constant. The set of all motif lengths for a code X and an organism s is

$$m^{(X,s)} = \left\{ m_{ij}^{(X,s)} | 1iN, 1jn_i \right\}$$

Again, we simply write m if the code and organism is known. m_k refers to an individual motif length ($1k|m|$) and $\bar{m}$ is the arithmetical mean of all lengths in m.

Next we show that the geometric distribution (with random variable G) can be used as a motif length distribution. Let us recall the geometric distribution: it has the distribution $P(G = l) = p(1-p)^l$ and shows how many times ($l = 1, 2, \ldots$) it takes until an event happens when the probability for that event is p. There is a critical assumption: the codons are uniformly distributed over the sequence according to their code usage. If the sequence contains any kind of pattern, e.g. repetitions, this assumption is questionable. On the other hand, what kind of pattern could we assume? We do not expect repetitions in coding sequences – other than in non-coding DNA. Let C represent the event where a codon of the code X is chosen and $N = \bar{C}$ where a codon not in X is chosen. The likelihoods are $P(C) = u$ (the code usage) and $P(\bar{C}) = 1 - u =: v$. If we assume a perfect sequence with uniform codon distribution, we get $u = 20/64$ and $v = 44/64$ (as the code X or any other maximal self-complementary circular code consists of 20 codons). The codon usage in biological sequences usually differs and so there will be different values for u and v.

Table 3
Example for motif lengths. C represents a codon of X and N a codon not part of X. The list of motifs is $m = (3, 1, 2)$.

codon sequence	NN	CCC	N	C	N	CC	N
motif lengths	2	3	1	1	1	2	1

Let us assume we encounter exactly one codon from X. The chance that the motif will end after one codon is v, since this is the likelihood that a non X code codon follows. Accordingly, a motif has the length of two with a chance of $u\cdot$ and so on. The length of a motif is the sequence where codons of X are read (events C) with a chance of u until the event $\bar{C}$ with a chance of v occurs. Thus, the probabilities need to be flipped in the geometric distribution, and with M as the random variable for motif lengths we get:

$$P(M = l) = \cdot(1-)^{l-1}$$

where $l = 1, 2, \ldots$ indicates the length of a motif.

The expectation value for a geometric distribution is $E(G) = 1/u$ and the variance $Var(G) = (1-u)/u^2$. The expected motif length for a given code usage u is then

$$E(M) = \frac{1}{-} = \frac{1}{1-u} \tag{1}$$

and the variance

$$Var(M) = \frac{1-}{^2} = \frac{u}{(1-u)^2}. \tag{2}$$

Therefore, once we know the code usage u we also know what average motif lengths to expect. Please note that the codon bias (see for example Athey et al., 2017) does affect the expected motif length $E(M)$ as the code usage u depends on the codon usage. If there are species with a low code usage we can expect shorter motif lengths. However, as we compare expected motif lengths with observed motif lengths based on the same code usage this does not matter. If we presume a geometric distribution, we can estimate a confidence interval and a p-value for the mean value of motif lengths to be greater than the expectation value. Using the estimator $\bar{M} = 1/n\sum_{k=1}^{n} m_k$ with n as the number of motif lengths and Eqs. (1) and (2),

$$Q = \frac{\bar{M} - \frac{1}{1-u}}{\sqrt{\frac{u}{n(1-u)^2}}}$$

will be asymptotically standard normal. This will later be used in Section 3.2.2.

3. Results and discussion

This sections presents the results. The code usage is briefly shown and then the motif lengths are discussed in more detail.

Table 4
Overview of code usage and motif lengths for eight model organisms. This table is continued on Table 5. These tables shows the first 20 codes out of 210 codes (column *Rank*) ranked and sorted by the code usage u (in percent). The remaining columns are the average motif lengths $\bar{m}$ and the difference in percent to the expected motif length ($\Delta\bar{m}$, in percent). Column p lists the p-values for average motif lengths to be greater than the expected mean motif length (99%). X_0 or C23 is highlighted in bold, C122 is marked with an *. Codes in *C. reinhardtii* ending with _2 are codes which are cyclic shifted two times.

Rank	*H. sapiens*	u (%)	$\bar{m}$	$\Delta\bar{m}$ (%)	p	*C. elegans*	u (%)	$\bar{m}$	$\Delta\bar{m}$ (%)	p	*S. cerevisiae*	u (%)	$\bar{m}$	$\Delta\bar{m}$ (%)	p
1	C25	44.6	1.82	0.6	0.0000	C122 *	40.4	1.75	4.4	0.0000	C122 *	40.5	1.77	5.0	0.0000
2	C117	44.1	1.82	1.4	0.0000	C24	40.2	1.71	2.5	0.0000	C171	40.5	1.76	5.0	0.0000
3	C173	44.1	1.79	0.1	0.1539	C171	40.0	1.71	2.8	0.0000	C20	39.8	1.72	3.7	0.0000
4	C98	43.7	1.79	0.7	0.0000	C97	39.9	1.71	2.6	0.0000	C123	39.7	1.72	3.6	0.0000
5	C166	43.7	1.80	1.4	0.0000	C25	39.7	1.70	2.6	0.0000	C21	39.7	1.72	4.0	0.0000
6	C20	43.3	1.77	0.4	0.0000	C26	39.7	1.71	2.9	0.0000	C22	39.1	1.69	2.6	0.0000
7	C111	42.8	1.76	0.8	0.0000	C20	39.6	1.70	2.7	0.0000	**C23** (X_0)	39.0	1.69	3.3	0.0000
8	**C23** (X_0)	42.8	1.75	−0.1	0.7767	C123	39.6	1.72	4.1	0.0000	C97	38.9	1.68	2.5	0.0000
9	C4	42.4	1.75	0.9	0.0000	C98	39.4	1.70	2.8	0.0000	C24	38.9	1.70	3.8	0.0000
10	C27	41.7	1.70	−0.7	1.0000	C172	39.3	1.70	3.0	0.0000	C168	38.7	1.71	4.8	0.0000
11	C24	41.1	1.68	−1.0	1.0000	C137	39.3	1.71	4.0	0.0000	C137	38.6	1.69	4.0	0.0000
12	C115	40.7	1.67	−0.9	1.0000	C27	39.2	1.68	2.3	0.0000	C161	38.4	1.68	3.5	0.0000
13	C172	40.7	1.65	−2.0	1.0000	C21	39.2	1.70	3.2	0.0000	C167	38.4	1.68	3.7	0.0000
14	C22	40.4	1.67	−0.5	1.0000	C22	39.1	1.70	3.4	0.0000	C110	38.3	1.67	3.0	0.0000
15	C97	40.2	1.65	−1.2	1.0000	C173	38.9	1.69	3.2	0.0000	C98	38.2	1.66	2.8	0.0000
16	C165	40.2	1.65	−1.3	1.0000	**C23** (X_0)	38.7	1.69	3.4	0.0000	C25	38.2	1.67	3.0	0.0000
17	C118	40.1	1.71	2.6	0.0000	C115	38.5	1.65	1.8	0.0000	C26	38.2	1.66	2.6	0.0000
18	C171	39.8	1.65	−0.9	1.0000	C161	38.3	1.65	1.9	0.0000	C172	38.1	1.67	3.1	0.0000
19	C76	39.6	1.70	2.6	0.0000	C117	38.0	1.65	2.4	0.0000	C120	37.9	1.67	3.5	0.0000
20	C161	39.4	1.63	−1.1	1.0000	C111	37.8	1.65	2.3	0.0000	C111	37.8	1.64	2.2	0.0000
Rank	*A. thaliana*	u (%)	$\bar{m}$	$\Delta\bar{m}$ (%)	p	*C. reinhardtii*	u (%)	$\bar{m}$	$\Delta\bar{m}$ (%)	p	*E. coli*	u (%)	$\bar{m}$	$\Delta\bar{m}$ (%)	p
1	C161	40.0	1.67	0.3	0.0014	C166	48.9	2.07	6.0	0.0000	**C23** (X_0)	46.9	1.92	2.1	0.0000
2	C171	39.9	1.68	1.3	0.0000	C173	48.2	2.02	4.7	0.0000	C20	46.5	1.92	3.0	0.0000
3	C115	39.4	1.66	0.5	0.0000	C117	48.1	2.03	5.4	0.0000	C22	45.8	1.88	1.9	0.0000
4	C24	39.3	1.66	0.9	0.0000	C25	47.4	1.98	4.2	0.0000	C173	45.7	1.86	1.1	0.0000
5	C3	39.1	1.65	0.7	0.0000	C98	47.1	1.97	3.9	0.0000	C25	45.3	1.87	2.1	0.0000
6	C21	39.0	1.66	1.2	0.0000	C4	46.1	1.98	6.8	0.0000	C27	44.6	1.81	0.4	0.0001
7	C111	38.7	1.66	1.5	0.0000	**C23** (X_0)	45.5	1.93	5.4	0.0000	C98	44.6	1.78	−1.4	1.0000
8	C107	38.7	1.64	0.3	0.0001	C111	45.3	1.95	6.4	0.0000	C4	43.6	1.79	1.2	0.0000
9	C122 *	38.6	1.68	2.9	0.0000	C27	44.8	1.89	4.3	0.0000	C111	43.1	1.79	1.7	0.0000
10	C20	38.6	1.66	1.8	0.0000	C20	44.6	1.90	5.2	0.0000	C166	42.4	1.75	0.6	0.0000
11	C165	38.5	1.64	0.9	0.0000	C188_2	44.0	1.89	6.0	0.0000	C117	41.9	1.74	1.3	0.0000
12	C172	38.4	1.64	0.8	0.0000	C165	44.0	1.94	8.9	0.0000	C21	41.9	1.88	9.1	0.0000
13	C110	38.3	1.61	−0.3	0.9999	C172	43.3	1.90	7.7	0.0000	C171	41.5	1.88	10.2	0.0000
14	C117	38.1	1.65	2.0	0.0000	C132_2	43.2	1.90	8.2	0.0000	C123	40.8	1.85	9.7	0.0000
15	C41	38.0	1.62	0.2	0.0307	C115	43.1	1.93	9.6	0.0000	C172	40.7	1.79	6.2	0.0000
16	C25	38.0	1.64	1.6	0.0000	C24	42.5	1.89	8.5	0.0000	C122 *	40.3	1.86	10.8	0.0000
17	C97	37.9	1.64	1.6	0.0000	C97	42.2	1.83	5.7	0.0000	C24	40.3	1.80	7.4	0.0000
18	C4	37.8	1.64	1.8	0.0000	C22	42.0	1.84	6.6	0.0000	C26	39.6	1.76	6.3	0.0000
19	C123	37.7	1.65	2.7	0.0000	C144_2	41.9	1.87	8.8	0.0000	C97	39.5	1.68	1.8	0.0000
20	**C23** (X_0)	37.7	1.63	1.6	0.0000	C3	41.2	1.94	14.1	0.0000	C109	39.1	1.62	−1.3	1.0000

3.1. Code usages

Tables 4 and 5 show the 20 best circular codes for frame 0 ranked by their code usage (column u) as well as their average motif lengths and the difference to the expected motif length which we will discuss later. More data is shown in the online supplementary material. X_0 (or C23) is always in the top 20 and has a code usage ranging from 46% in *C. reinhardtii* (rank 7) to 39% in *S. cerevisiae* (rank 7). In *E. coli* X_0 is on position 1 with $u = 47\%$. Similar figures are valid for the code usages in frame 1 and frame 2 where the tuple-wise cyclic codes were dominant (data not shown, see online supplementary material). The code usages of X_0 in different species are definitely higher than expected for circular codes. One would expect 20/64 or about 31%. Thus, the results of Arquès and Michel (1996) and Michel (2015) could be confirmed. Not only X_0 has a good code coverage. Also some other circular codes with similar properties (in $\mathcal{C}$) are significant. The random codes ($\mathcal{R}$) show a maximal code usage of about 32–45% for different species in frame 0 and slightly smaller values for frame 1 and 2 (see online supplementary material).

3.2. Motif lengths

3.2.1. Motif lengths in relationship to code usage

Let us start with the results of random sequences. Fig. 2a) shows the code usage (x-axis) plotted against the average motif length (y-axis). Filled blue samples indicate the 210 circular codes ($\mathcal{C}$) and red diamonds show 210 random codes ($\mathcal{R}$). Evidently, the motif lengths of circular codes are not decisively longer than any random code (of 20 codons). Some of the circular codes even have shorter motif lengths than expected. The code usage of the circular codes is close to 31% (20/64) whereas the code usage of the random codes has a higher variance. The dots lie on the line that indicates the expectation value $E(M)$ (see Eq. (1)). Fig. 2b) shows the situation in the DNA section of human chromosome 1 which is quite similar with one exception: many of the random codes have a slightly higher average motif length than expected. We think this is because of repetitive patterns in the DNA sequence. In particular, the random codes may have codons like AAA that consist of the same nucleotide. If a sequence has repetitions of such a nucleotide, the average motif length will increase.

The biological coding sequences show a slightly different picture. The average motif lengths were in all organisms surprisingly short, too. As seen in Table 4 (column $\bar{m}$) they lie in a range of 1.8 (*C. elegans*) to 2.1 (*C. reinhardtii*). X_0 is – like in the case of code usage – not the best code per species codons and its average motif length is slightly shorter than the best in the respective organism.

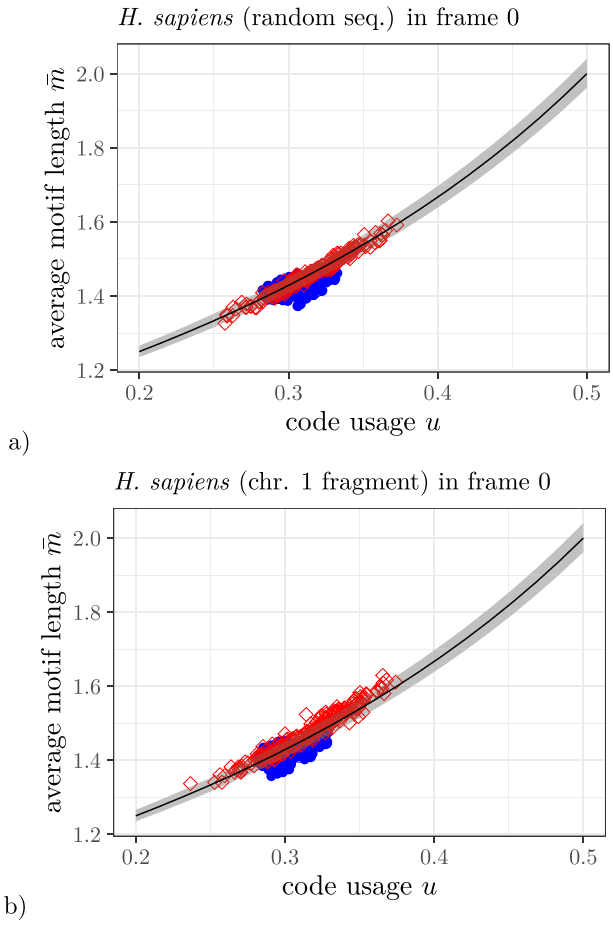

Fig. 2. Code usage u and average motif lengths $\bar{m}$ in a) artificial sequence and b) a region of human chromosome 1 (both 100,000 bases long). Filled blue samples (●) indicate the 210 circular codes ($\mathcal{C}$) and red diamonds (◇) show 210 random codes ($\mathcal{R}$). The black line (／) shows the expectation value $E(M)$. The gray area indicates the confidence interval of $E(M)$ ($\alpha = 0.01$).

Table 5
Table 4 continued.

Rank	*P. occultum*	u (%)	$\bar{m}$	$\Delta\bar{m}$ (%)	p	*T. tenax*	u (%)	$\bar{m}$	$\Delta\bar{m}$ (%)	p
1	C166	52.7	2.13	0.7	0.0007	C166	47.5	1.93	1.5	0.0000
2	C165	51.4	2.03	−1.1	1.0000	C98	46.8	1.93	2.6	0.0000
3	C173	51.2	2.07	0.9	0.0000	C173	46.6	1.92	2.7	0.0000
4	C172	49.8	1.98	−0.6	0.9988	C165	46.6	1.92	2.7	0.0000
5	C4	49.4	1.99	0.5	0.0133	C97	45.9	1.93	4.3	0.0000
6	C117	49.1	1.96	−0.3	0.9030	C172	45.7	1.92	4.3	0.0000
7	C98	49.1	1.98	0.8	0.0001	C117	44.9	1.81	−0.5	0.9986
8	C3	48.1	1.91	−0.9	1.0000	C4	44.6	1.84	1.9	0.0000
9	**C23** (X_0)	47.9	1.93	0.6	0.0015	C25	44.0	1.79	0.2	0.1322
10	C115	47.8	1.87	−2.3	1.0000	C115	44.0	1.78	−0.2	0.8898
11	C97	47.7	1.90	−0.9	1.0000	**C23** (X_0)	43.7	1.83	3.0	0.0000
12	C25	47.6	1.90	−0.4	0.9819	C3	43.7	1.82	2.2	0.0000
13	C21	46.5	1.86	−0.5	0.9947	C24	43.1	1.77	0.9	0.0000
14	C24	46.2	1.82	−2.2	1.0000	C21	42.8	1.81	3.6	0.0000
15	C111	45.8	1.83	−0.8	1.0000	C132_2	42.3	1.74	0.4	0.0097
16	C161	44.5	1.76	−2.5	1.0000	C111	42.0	1.72	−0.3	0.9388
17	C20	44.3	1.77	−1.1	1.0000	C188_2	41.4	1.70	−0.3	0.9734
18	C171	42.9	1.71	−2.5	1.0000	C20	41.1	1.70	0.4	0.0184
19	C118	42.8	1.81	3.2	0.0000	C161	41.1	1.68	−0.9	1.0000
20	C41	42.7	1.70	−2.6	1.0000	C27	40.6	1.73	3.0	0.0000

The differences (in percent) relative to the expectation value of the motif lengths ($E(M)$, see Eq. (1)) are not very high within the best codes (compare Table 4) but they are visible. Some average motif lengths in *H. sapiens* are even smaller for codes with good code coverage. X_0 is an example ($\Delta\bar{m} = -0.1$ %). There are, however, motifs which are rather long: In *S. cerevisiae* code X_0 has motifs up to a length of 210 codons. In *H. sapiens* and *C. elegans* the maximal motif lengths are 44 and 39 codons, respectively, for code $C122$ (data not shown). This, again, confirms the finding in El Soufi and Michel (2016) where long motifs in coding sequences were identified.

Fig. 4 shows as an example the code usage plotted against the motif lengths in *S. cerevisiae* for frame 0 and 1. The best 20 circular codes clearly have a greater average motif length than the best random codes have – the curve has a sickle shape. This is true for frame 0 (cf Fig. 4a) but surprisingly also for frame 1 (cf Fig. 4b) and frame 2 (see online supplementary material). X_0 (or $C23$) is on position 7 and has a 3% higher average motif length than expected (see Table 4). The winner is a self-complementary circular code labeled as $C122$ which is also in position 1 in *C. elegans*. This code has the highest code usage and the highest increase of the average motif length relative to the expectation value. $C122$ is in the top 20 in Table 4 except for green alga *C. reinhardtii* and

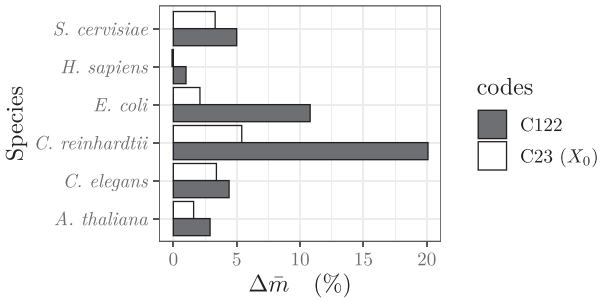

Fig. 3. Differences $\Delta\bar{m}$ in % for average motif length relative to expected motif length for codes X_0 (white) and $C122$ (gray).

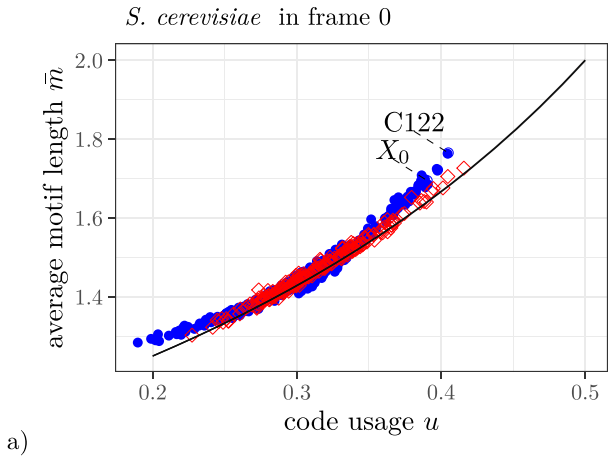

Fig. 4. Code usage u and average motif lengths $\bar{m}$ for *S. cerevisiae* in frame 0 (a) and frame 1 (b). Filled blue samples (●) indicate the 210 circular codes ($\mathcal{C}$) and red diamonds (◇) show 210 random codes ($\mathcal{R}$). The black line (／) shows the expectation value $E(M)$.

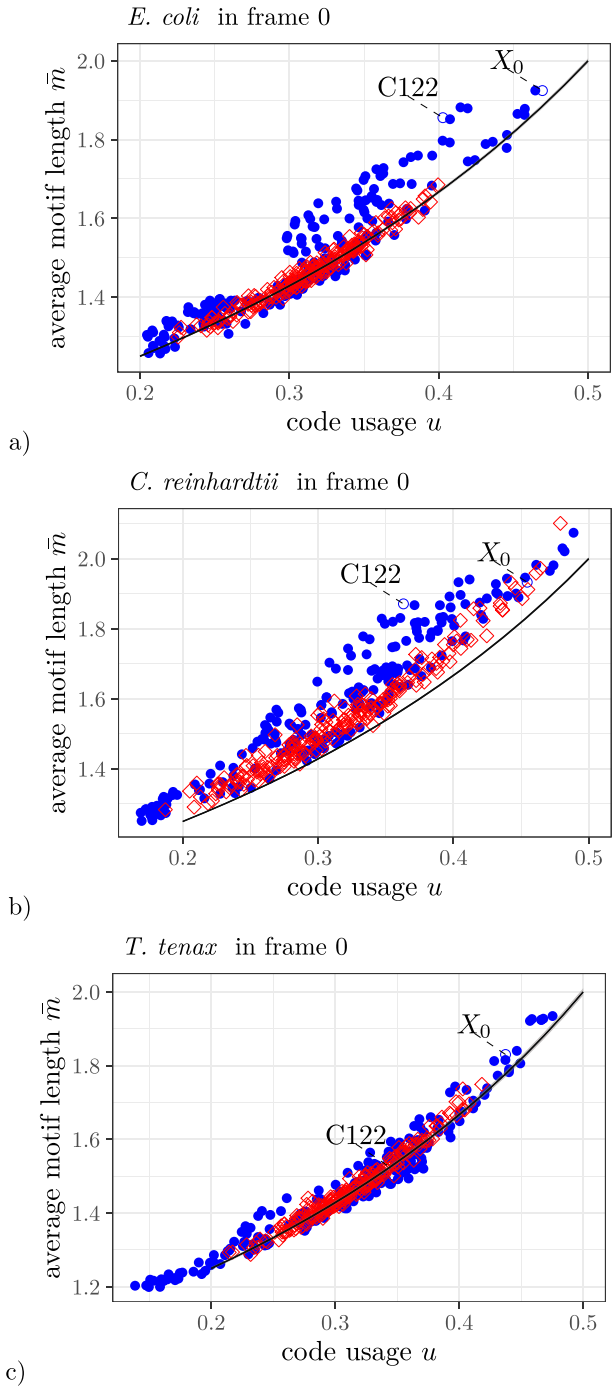

Fig. 5. Code usage u and average motif lengths $\bar{m}$ for *E. coli*, *C. reinhardtii* and *T. tenax*. Filled blue samples (●) indicate the 210 circular codes ($\mathcal{C}$) and red diamonds (◇) show 210 random codes ($\mathcal{R}$). The black line (／) shows the expectation value $E(M)$.

the two archaea. The alga's code usage is not very high but the difference to the expectation value $\Delta\bar{m}$ is about 20% (details see online supplementary material). Fig. 3 shows all organisms in an overview. $C122$ contains these codons:

$$C122 = \{\text{AAC}, \text{AAT}, \text{ACC}, \text{ACT}, \text{AGC}, \text{AGT}, \text{ATC}, \text{ATT}, \text{GAA}, \text{GAC}, \text{GAT}, \text{GCC}, \text{GCT}, \text{GGA}, \text{GGC}, \text{GGT}, \text{GTC}, \text{GTT}, \text{TCC}, \text{TTC}\}$$

The differences between X_0 (or $C23$) and $C122$ are six codons (out of 20):

- CAG, CTC, TAC, CTG, GAG, GTA only in $C23 = X_0$
- AGC, TCC, ACT, GCT, GGA, AGT only in $C122$

All different codons are cyclic shifted versions: Three of them are shifted for one and the other three for two positions. This is because of the way maximal self-complementary codes are constructed.

Besides *S. cerevisiae*, the organisms *E. coli* and *C. reinhardtii* show interesting patterns as depicted in Fig. 5. *E. coli* shows the most striking increase of the average motif lengths relative to random codes. Again code $C122$ has the highest increase of about 11% whereas X_0 only has about 2% though this code is in position 1 according to the code usage u (see Table 4). *C. reinhardtii* shows an exceptional pattern: many circular codes including X_0 have a higher average motif length than expected but also random codes show higher than expected motif lengths in this organism – although less explicit. This suggests that coding sequences of *C. reinhardtii* have many repetitive regions, similar to plain DNA found on chromosomes (as shown in Fig. 2b).

We observed that many circular codes in general have longer motifs than random codes in coding sequences. Fig. 6 shows an example for *E. coli*. A similar pattern could be observed in all eight model organisms with the weakest effect in *H. sapiens* and the strongest in *E. coli* (as in Fig. 6) and *C. reinhardtii* (see also Fig. 5). Overall, all circular codes in all eight model organisms show a positive difference to the expectation value of in average 2.4%. Surprisingly, also the random codes have a positive difference of 1.2%, on average. Again, we think that this is caused by repetitive patterns in the coding sequences (see also above). Note that there are also circular codes which have motif lengths below average. This is to some extend to be expected, as the 210 circular codes contain 140 tuple-wise shifted codes which will not perform well in frame 0. With respect to recognizing frame shifts, it is nevertheless sufficient that there is at least one circular code with long motifs.

3.2.2. Test for geometric distribution and statistical significance

The motif lengths m were tested by a χ^2 test to see if they follow a geometric distribution. The null-hypothesis H_0 is: all motif lengths are distributed according to the geometric distribution. The p-values for every code and every organism are very close to 1 indicating a strong relationship to a geometric distribution – at least H_0 cannot be rejected (see online supplementary material for details). Thus, we can assume that the motif lengths are geometrically distributed. Fig. 7 shows the histogram of motif lengths for *E. coli* in frame 0 for code $C122$. In relation to the statistical motif length distribution it has less motifs of length 1 and more of length 2 or greater.

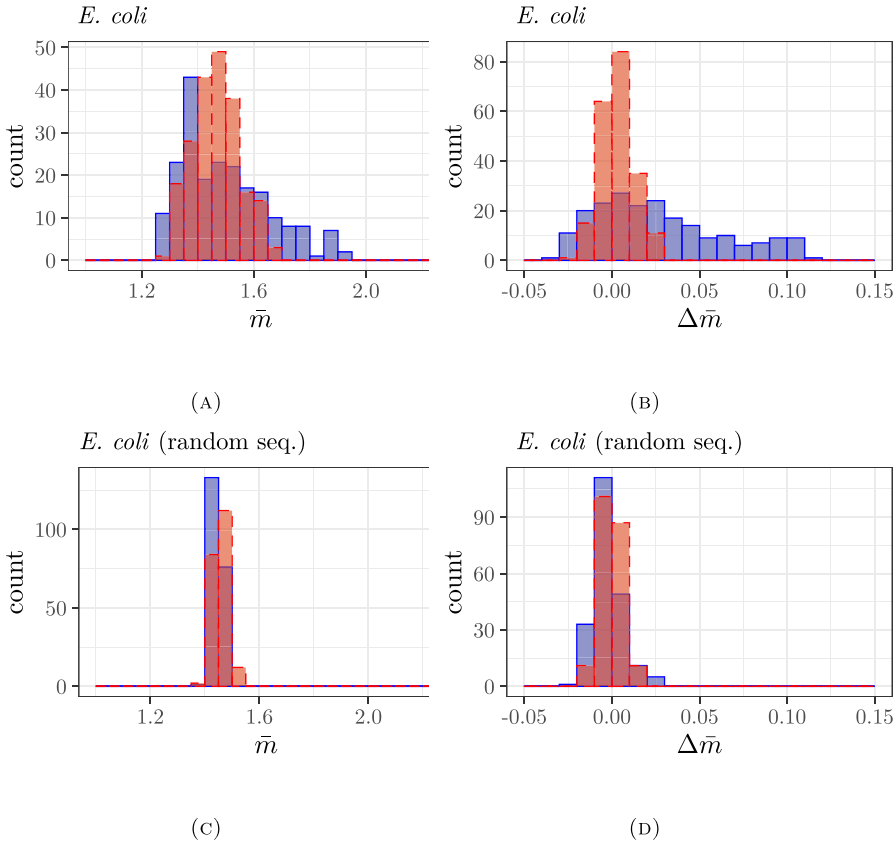

Fig. 6. Motif length histograms in *E. coli* for circular (transparent blue) and random codes (transparent red). A ruby color shows overlapping bars. Left column: a) Average motif lengths histogram. c) Average motif length histogram in a *E. coli*-like random sequence. Right column: b) Histogram of motif length differences relative to the expectation value ($\Delta\bar{m}$). d) Same for a *E. coli*-like random sequence. Many circular codes exceed the motif lengths of random codes in *E. coli* coding sequences (a). The effect is even stronger when the differences to the expected motif length ($\Delta\bar{m}$) are considered (b). There is no such effect for random sequences, neither for motif lengths (c) nor for differences to the expectation value (d).

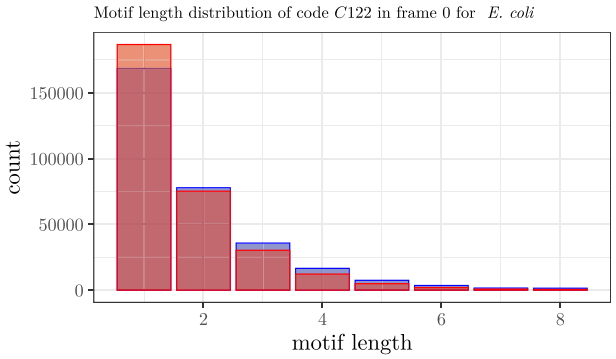

Fig. 7. Motif lengths distribution in *E. coli* in frame 0. Transparent blue bars indicate circular code $C122$, transparent red bars represent the theoretical geometric distribution with the same code usage u that $C122$ has (40.3 %). A ruby color shows overlapping bars. All motif lengths greater than or equal to 8 are summarized in one bar (labeled as 8). Evidently, $C122$ has less motifs of length 1 and more of length 2 or greater.

If we can presume a geometric distribution and starting from Eq. (2.3) we can estimate a confidence interval and a p-value for the mean value of motif lengths being greater than $E(M)$. The sample size, i.e. the number of motif lengths $|m|$ per organism, is very high (about 300,000). The confidence interval for the estimator $\overline{M}$ ($\alpha = 0.01$) is very close to the expectation value and almost invisible in Figs. 2, 4, and 5. Also, many p-values for $P(\overline{M} > E(M))$ in Table 4 (column p) are close to 0. This indicates that the average lengths of circular codes motifs are statistically significant longer than expected by chance.

4. Conclusions

Circular codes are over-represented in coding regions. It is an open question why these specific codons are used. As we have shown, there is little evidence that circular codes are used for frame shift recognition throughout the total length of a coding sequence. However, there are regions in transcribed sequences which possess very long motifs, although there are not many of them in all the sequences assessed. Also, the average motif lengths are significantly higher than expected for many codes, though the difference ($\Delta\bar{m}$) is not large enough to imply long motifs in all coding sequences. X_0 is the best average circular code over all species tested here. Our analysis suggests that other maximal self-complementary codes like C122 might be the *best* code for a specific organism. C122 is a variation of X_0 and differs only in six codons. In particular, in comparatively older organisms like *C. reinhardtii* and comparatively simpler organisms like *E. coli* the motif lengths are, on average, much longer than would be expected. Nevertheless, this was not the case in the two archaea.

We have analyzed eight model organisms. Although the overall course of the scatter plots (motif lengths over code usage) is similar, there are distinct differences between the species. Surprisingly, the coding sequences in the green alga *C. reinhardtii* seems to have many repetitive sequences, even in coding regions.

As we have shown, the average or expected motif lengths depend on the code usage. If, however, the code usage was fixed but the codons were resorted to build clusters of motifs, then the motif lengths can be increased. Our work showed that random codes can lead to long motifs on average, as well. As these results do not differ much from those of circular codes, a more comprehensive statistical analysis should be performed in this respect.

Lastly, if we assume that a biological entity like the ribosome could, while progressing on the mRNA, distinguish codons of a circular code from other codons, one important point needs to be addressed: as long as there is no frame shift and the ribosome reads through a motif, there is no issue. If, nevertheless, the ribosome encounters a stretch of codons *not* from a circular code, there is an ambiguous situation: either there was a frame shift and translation should be aborted or the ribosome just reads a non-motif region and translation should be continued. Since in this study we always found coding sequence with gaps of non-motifs between circular code motifs, we speculate that frame shift prevention by means of circular codes is not an important mechanism in present day organisms. Since, on the other hand, we do find above average motifs in coding sequences, our assumption is that circular codes could have been a means to control frame shifts in organisms more ancient than the ones we assessed.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Credit authorship contribution statement

M. Gumbel: Conceptualization, Methodology, Software, Visualization, Writing - original draft. **P. Wiedemann:** Investigation, Data curation, Writing - original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Thomas Ihme for the use of the CUDA server to run their analysis and John Clear for critical comments on the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jtbi.2021.110708.

References

Arquès, Didier G., Michel, Christian J., 1996. A complementary circular code in the protein coding genes. Journal of Theoretical Biology 182 (1), 45–58.

Athey, John, Alexaki, Aikaterini, Osipova, Ekaterina, Rostovtsev, Alexandre, Santana-Quintero, Luis V., Katneni, Upendra, Simonyan, Vahan, Kimchi-Sarfaty, Chava, 2017. A new and updated resource for codon usage tables. BMC Bioinformatics 18 (391).

Crick, F.H., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. Proceedings of the National Academy of Sciences of the United States of America 43, 416–421..

Dila, Gopal, Michel, Christian J., Poch, Olivier, Ripp, Raymond, Thompson, Julie D., 2019. Evolutionary conservation and functional implications of circular code motifs in eukaryotic genomes. Bio Systems 175, 57–74..

Drummond, Allan D., Wilke, Claus O., 2009. The evolutionary consequences of erroneous protein synthesis. Nature Reviews. Genetics 10, 715–724.

Dunkle, Jack A., Dunham, Christine M., 2015 Mechanisms of mrna frame maintenance and its subversion during translation of the genetic code. Biochimie 114, 90–96..

El Soufi, Karim, Michel, Christian J., 2016. Circular code motifs in genomes of eukaryotes. Journal of Theoretical Biology 408, 198–212.

Fijalkowska, Iwona J., Schaaper, Roel M., Jonczyk, Piotr, 2012. Dna replication fidelity in escherichia coli: a multi-dna polymerase affair. FEMS Microbiology Reviews 36, 1105–1121.

Fimmel, Elena, Giannerini, Simone, Gonzalez, Diego Luis, Strüngmann, Lutz, 2015. Circular codes, symmetries and transformations. Journal of Mathematical Biology 70 (7), 1623–1644..

Fimmel, Elena, Gumbel, Markus, Karpuzoglu, Ali, Petoukhov, Sergey, April 2019. On comparing composition principles of long dna sequences with those of random ones. Bio Systems 180, 101–108.

Fimmel, Elena, Strüngmann, Lutz, 2018. Mathematical fundamentals for the noise immunity of the genetic code. Biosystems 164, 186–198.
Gamper, H.B., Masuda, I., Frenkel-Morgenstern, M., Hou, Y.M., 2015. Maintenance of protein synthesis reading frame by ef-p and m(1)g37-trna. Nature Communications 6 (7226)..
Meyerovich, Mor, Mamou, Gideon, Ben-Yehuda, Sigal, 2010. Visualizing high error levels during gene expression in living bacterial cells. Proceedings of the National Academy of Sciences of the United States of America 107, 11543–11548..
Michel, Christian J., 2012. Circular code motifs in transfer and 16s ribosomal rnas: a possible translation code in genes. Computational Biology and Chemistry 37, 24–37..
Michel, Christian J., 2014. A genetic scale of reading frame coding. Journal of Theoretical Biology 355, 83–94..
Michel, Christian J., Pirillo, Giuseppe, Pirillo, Mario A., 2008. A relation between trinucleotide comma-free codes and trinucleotide circular codes. Theoretical Computer Science 401, 17–26.
Michel, C.J., 2015. The maximal c3 self-complementary trinucleotide circular code x in genes of bacteria, eukaryotes, plasmids and viruses. Journal of Theoretical Biology 380, 156–177.
Warnecke, Tobias, Huang, Yang, Przytycka, Teresa M., Hurst, Laurence D., 2010. Unique cost dynamics elucidate the role of frameshifting errors in promoting translational robustness. Genome Biology and Evolution 2, 636–645.