Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

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

A large-scale analysis of codon usage bias in 4868 bacterial genomes shows association of codon adaptation index with GC content, protein functional domains and bacterial phenotypes

Anna Masłowska-Górnicz, Melanie R.M. van den Bosch, Edoardo Saccenti *, Maria Suarez-Diez *

Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Stippeneng 4, 6708 WE Wageningen, the Netherlands

ARTICLE INFO

Keywords:
Bacterial genome
Codon adaptiveness
Evolution

ABSTRACT

Multiple synonymous codons code for the same amino acid, resulting in the degeneracy of the genetic code and in the preferred used of some codons called codon bias usage (CBU). We performed a large-scale analysis of codon usage bias analysing the distribution of the codon adaptation index (CAI) and the codon relative adaptiveness index (RA) in 4868 bacterial genomes. We found that CAI values differ significantly between protein functional domains and part of the protein outside domains and show how CAI, GC content and preferred usage of polymerase III alpha subunits are related. Additionally, we give evidence of the association between CAI and bacterial phenotypes.

1. Introduction

The 20 amino acids required for protein synthesis are encoded by 61 codons: this implies a redundancy in the genetic code, as several codons, so-called synonymous, encode for the same amino acid (AA). Synonymous codons do not alter the encoded AA sequence and were considered to be equivalent and interchangeable [1]. However, it has been observed that certain sets of codons are preferred over others. This phenomenon, which happens across all domains of life, is called codon usage bias (CUB), and has been shown to affect many different biological processes [2]. For instance, preference for certain codons correlates with high protein expression [3]: this has been explained by differences in the relative abundance of the corresponding transfer RNAs (tRNAs) which determines the efficiency and accuracy of protein translation [4,5]. Codons that are highly adapted to the tRNA pool are recognized by abundant tRNAs [6–8] and transcripts with codons biased towards the more abundant tRNAs are often found to have higher translation rates.

The codon adaptation index (CAI) has been proposed as a measure to quantify the frequency of preferred or optimal codons in a given gene [9], and it is based on the usage frequency of a given codon i in a reference set of genes coding for highly abundant proteins. This frequency is termed relative adaptiveness $RA_{AA,i}$ of codon i associated to amino acid AA. Codons with high RA lead to higher CAI values and it is assumed that codons associated to genes coding for highly abundant proteins (codons with high RA) are the codons that are recognized by the most abundant and efficient tRNA species in an organism [10,11]. In addition to codon usage bias, other preference variations have been observed such as codon pairs which are associated with translation elongation rate and protein folding efficiency suggesting that optimality of individual codons and properties of adjacent codon pairs both contribute to gene regulation [12,13].

GC content variation in bacterial genomes has also been associated with CUB, as a result of the presence/absence of polymerase III alpha sub-unit isoforms and distinct groups of bacterial genomes, with specific spectra of GC content variation, have been observed [14,15] and associated with the dimeric combination of sub-units: dnaE1 (homodimer, full-spectrum), dnaE2/dnaE1 (heterodimer, high-GC) and polC/dnaE3 (heterodimer, low-GC) [14]. This suggests an indirect link between CUB and the preferred use of a certain type of polymerase in a genome [14,15].

Differences in CUB and associated CAI measures have also been related to structural differences in the coded protein. Structural domains (SDs) are the protein parts with a defined tertiary structure which provides functional properties; in contrast, intrinsically disordered regions (IDRs) are stretches of amino acids that are either unfolded in solution or show non-globular structures of undefined confirmation and are free from structural constraints. CUB has been observed to characterize both SDs and IDRs in eukaryotes: gene segments encoding IDRs tend to have a

* Corresponding authors.
E-mail addresses: edoardo.saccenti@wur.nl (E. Saccenti), maria.suarezdiez@wur.nl (M. Suarez-Diez).

https://doi.org/10.1016/j.bbagrm.2022.194826
Received 10 January 2022; Received in revised form 5 May 2022; Accepted 12 May 2022
Available online 20 May 2022

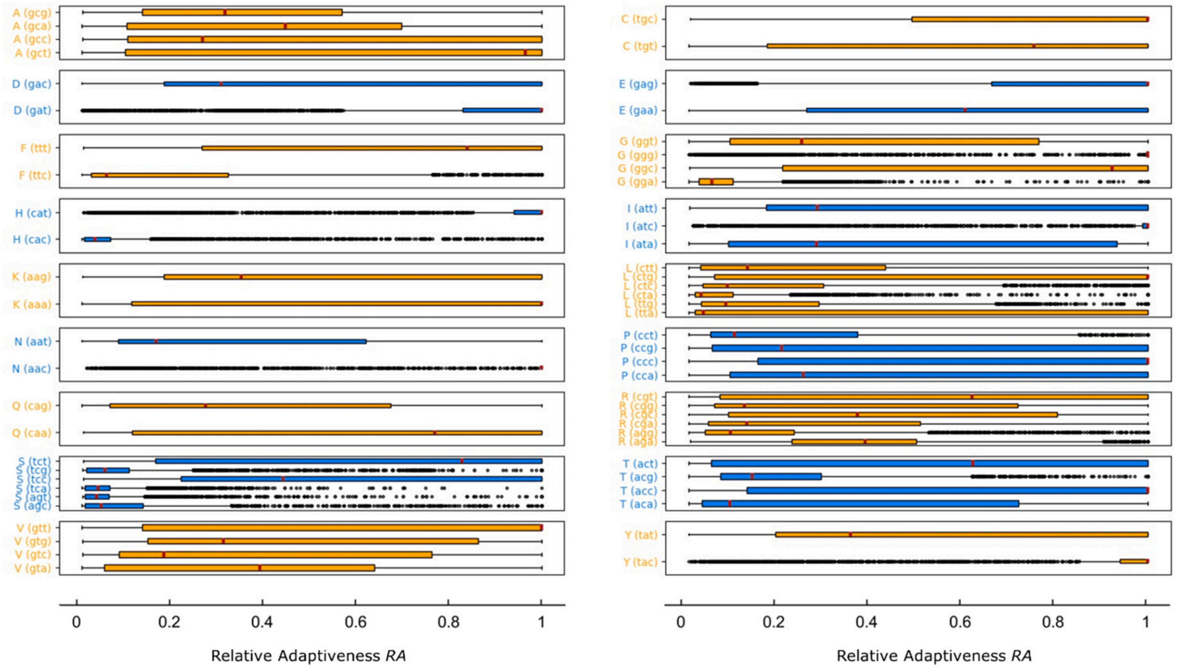


Fig. 1. Distribution of codon relative adaptiveness *RA*, Eq. (1), across 4868 bacterial genomes per amino-acids RA is calculated using the modified OPTIMIZER algorithm [33]. Amino acid one-letter code. A alanine; C cysteine; D aspartic acid; E glutamic acid; F phenylalanine; G glycine; H histidine; I isoleucine; K lysine; L leucine; M methionine; N asparagine; P proline; Q glutamine; R arginine; S serine; T threonine; V valine; W tryptophan; Y tyrosine. Colours are used to enhance readability of the plots [18].

lower tRNA adaptation index (*tAI*) than those corresponding to SDs. The tAI provides an alternative measurement of CUB [16], indicating that IDRs have less optimized codon usage than SDs [17]. The preferential use of non-optimal codons in genome regions predicted to be intrinsically disordered has been observed in the filamentous fungus *Neurospora* [18]. Protein domain boundaries are enriched in optimal codons but no evidence of non-optimal codons clustered around domain boundaries in *Escherichia coli*, *Saccharomyces cerevisiae* or *Homo sapiens* has been found [19], suggesting translational pausing at domain boundaries to be incorporated in the nucleotide sequence to promote folding accuracy.

In this work we investigate the CUB and the distribution of CAI in 4868 bacterial genomes as well as the distribution of preferred codons within and across taxonomic levels and their relationship with protein structure and genome characteristics. Consistently with actual knowledge we observed that nucleotide usage at third position may have significant impact on the bacterial genome characteristics. We found that CAI values differ significantly between protein functional domains and part of the protein outside domains, as a result of CUB, and show how CAI, GC content and usage of polymerase III alpha subunits are related. Additional we give evidence of the association between CAI and bacterial phenotypes.

2. Materials and methods

2.1. Data retrieval and genome annotation

4868 bacterial genomes were obtained from the European Nucleotide Archive (ENA) in EMBL format using enaBrowserTools (https://github.com/enasequence/enaBrowserTools). The genomes were consistently re-annotated using the Semantic Annotation Platform with Provenance (SAPP) [20]. Within the SAPP framework, *de-novo* gene prediction was performed using Prodigal 2.6.2 [21] and protein coding sequences were annotated using InterProScan 5.4-47.0 [22] (modules selected were: TIGRFAM [22], PIRSF [23], ProDom [24], SMART 5 [25], PROSITE [26], HAMAP [27], Pfam [28], PRINTS [29], SUPERFAMILY [30] and Gene3D [31]).

The *de novo* annotated genomes and corresponding annotation provenance were stored in RDF (Resource Description Format) according to the GBOL [32] ontology in a triplestore graph-database (Blazegraph Workbench v2.1.0) for further analysis. In the subsequent analysis the domain annotation used was the one performed using Pfam. For a list of genomes analysed see *final_weight_table_with_taxonomy.csv* available as Supplementary material.

2.2. Codon relative adaptiveness

The relative adaptiveness $RA_{\mathrm{AA},i}(G)$ for the *i*-th codon associated to amino acid AA, given a set *G* of genes is defined as [11]:

$$RA_{AA,i}(G) = \frac{f_{AA,i}}{f_{j,max}} \tag{1}$$

where $f_{AA,i}$ is the frequency of the *i*-th codon for the AA amino acid and $f_{AA,\mathrm{max}}$ is the maximum value of the frequency when considering all synonymous codons associated to AA. The calculation of $RA_{\mathrm{AA},i}(G)$ is discussed in Section 2.4.

2.3. Codon adaptation index

The codon adaptation index *CAI*(*G*) for a set *G* of genes is defined as

$$CAI(G) = \left(\prod_{i=1}^{N} RA_{AA,i}(G)\right)^{\frac{1}{N}} \tag{2}$$

CAI values ranges from 0 to 1: values close to 1 result from usage of codons with high frequency (a large number of optimal codons) in the *G* set of genes.

2.4. Calculation of relative adaptiveness

Calculation of the relative adaptiveness $RA_{\mathrm{AA},i}$ of each codon for a genome requires the identification of a set *G* of genes (see Eq. (1)) corresponding to highly abundant proteins. This is usually achieved by

Table 1
Descriptive statistics for the relative adaptiveness (*RA*, Eq. (1)) of 59 codons calculated over 4868 bacterial genomes. RA is calculated using the modified OPTIMIZER algorithm (17) as shown in Fig. 1.

Amino acid	Codon	Mean	Standard deviation	Median	MAD	Min
A	(gct)	0.65	0.42	0.97	0.05	0.001
A	(gcc)	0.46	0.4	0.26	0.32	0.002
A	(gca)	0.45	0.33	0.44	0.45	0.001
A	(gcg)	0.38	0.29	0.31	0.29	0.002
C	(tgt)	0.59	0.4	0.75	0.36	0.007
C	(tgc)	0.76	0.32	1	0	0.012
D	(gat)	0.85	0.28	1	0	0.001
D	(gac)	0.46	0.36	0.3	0.29	0.002
E	(gaa)	0.61	0.36	0.6	0.59	0.002
E	(gag)	0.79	0.33	1	0	0.006
F	(ttc)	0.23	0.31	0.05	0.07	0.001
F	(ttt)	0.66	0.37	0.84	0.24	0.003
G	(gga)	0.09	0.16	0.05	0.05	0.001
G	(ggc)	0.65	0.39	0.92	0.11	0.003
G	(ggg)	0.85	0.31	1	0	0.002
G	(ggt)	0.41	0.37	0.25	0.3	0.002
H	(cac)	0.11	0.23	0.03	0.04	0.001
H	(cat)	0.84	0.31	1	0	0.004
I	(ata)	0.44	0.38	0.28	0.33	0.002
I	(atc)	0.85	0.29			0.011
I	(att)	0.48	0.37	0.28	0.28	0.004
K	(aaa)	0.66	0.43	1	0	0.002
K	(aag)	0.54	0.4	0.35	0.43	0.002
L	(tta)	0.35	0.44	0.03	0.05	0.001
L	(ttg)	0.2	0.26	0.08	0.09	0.001
L	(cta)	0.09	0.16	0.03	0.04	0.001
L	(ctc)	0.23	0.29	0.08	0.09	0.001
L	(ctg)	0.59	0.45	1	0	0.001
L	(ctt)	0.28	0.33	0.13	0.17	0.001
N	(aac)	0.87	0.29	1	0	0.012
N	(aat)	0.36	0.37	0.16	0.17	0.003
P	(cca)	0.52	0.45	0.25	0.36	0.002
P	(ccc)	0.64	0.42	1	0	0.003
P	(ccg)	0.44	0.43	0.2	0.29	0.002
P	(cct)	0.24	0.29	0.1	0.13	0.003
Q	(caa)	0.59	0.43	0.77	0.34	0.005
Q	(cag)	0.38	0.34	0.27	0.34	0.002
R	(aga)	0.42	0.26	0.39	0.22	0.005
R	(agg)	0.19	0.25	0.09	0.1	0.002
R	(cga)	0.32	0.36	0.13	0.18	0.002
R	(cgc)	0.46	0.37	0.37	0.47	0.003
R	(cgg)	0.34	0.39	0.12	0.16	0.003
R	(cgt)	0.56	0.42	0.62	0.57	0.002
S	(agc)	0.21	0.35	0.04	0.06	0.001
S	(agt)	0.06	0.12	0.03	0.04	0.002
S	(tca)	0.05	0.08	0.04	0.04	0.002
S	(tcc)	0.56	0.38	0.44	0.54	0.004
S	(tcg)	0.11	0.18	0.05	0.06	0.002
S	(tct)	0.61	0.41	0.83	0.25	0.004
T	(aca)	0.35	0.38	0.09	0.13	0.002
T	(acc)	0.62	0.43	1	0	0.002
T	(acg)	0.23	0.25	0.14	0.13	0.002
T	(act)	0.52	0.41	0.62	0.56	0.002
V	(gta)	0.39	0.32	0.39	0.47	0.002
V	(gtc)	0.38	0.37	0.18	0.2	0.003
V	(gtg)	0.46	0.36	0.31	0.32	0.002
V	(gtt)	0.67	0.42	1	0	0.002
Y	(tac)	0.84	0.3	1	0	0.014
Y	(tat)	0.5	0.36	0.35	0.37	0.003

complementing genome information with experimental measurements. Here we have used the conceptual framework proposed by Puigbo et al. [33] and implemented in the OPTIMIZER web-server (http://genomes.urv.es/OPTIMIZER/). This is an on-line application where codon usage is used to optimize DNA sequences for expression in a different host. The OPTIMIZER uses relative abundance tables for 150 organisms that have been computed using an iterative algorithm (see Fig. S1 in the Supplementary material) that starts with the initial selection of a set of 25 genes encoding for ribosomal proteins, as ribosomal proteins are assumed to be among those with the highest expression. Afterwards, iterative calculations of RA (see Eq. (1)) followed by *CAI* scoring (see Eq. (2)) of all genes in the genome and subsequent selection of genes with highest *CAI* for RA calculation is performed until convergence, that is until the selected set of genes remains stable between iterations.

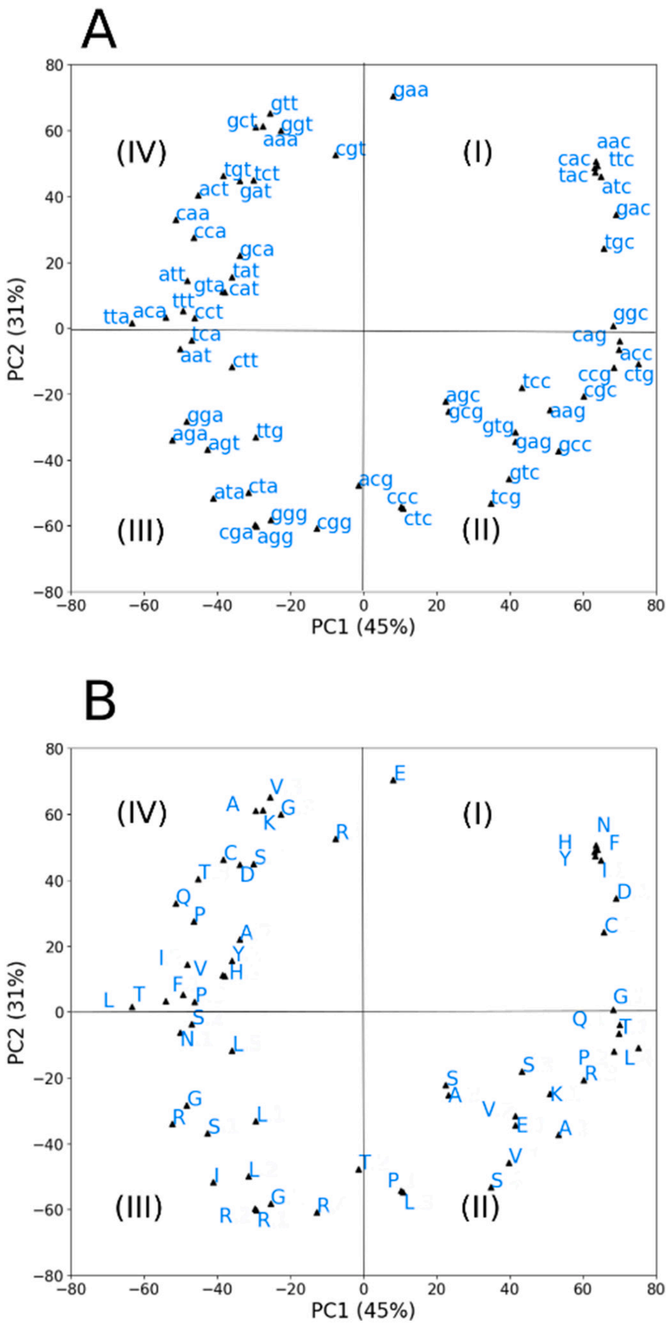


Fig. 2. Score plot of a principal component analysis on the 59 × 4868 matrix containing the relative adaptiveness for each codon specific to each genome labelled by A) codon B). See caption of Fig. 1 for AA letters code legend.

To extend the OPTIMIZER calculations beyond the original selection 150 organism we have re-coded the algorithm. We also modified the original OPTIMIZER algorithm by implementing a random selection of the initial set of bacterial genes to reduce the dependency of relative adaptiveness calculations on gene functional annotation information. This implementation gives RA values similar to those obtained using the original implementation. Testing results are given Section S1 and Fig. S2 of the Supplementary material.

To verify further the suitability of this approach we compared the

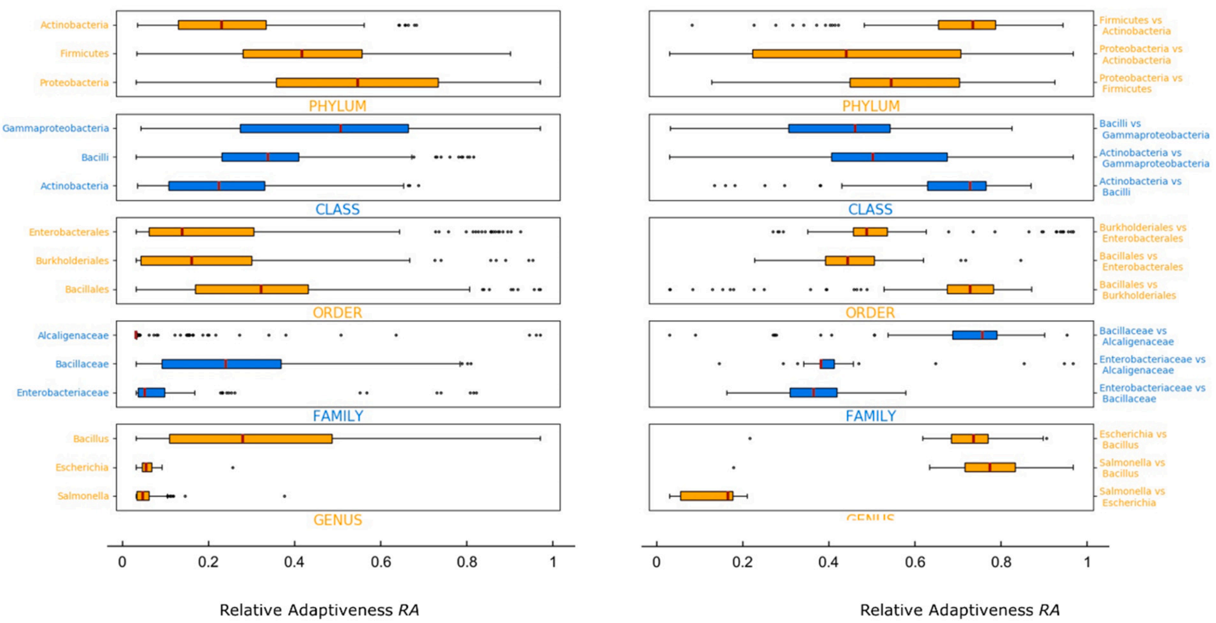


Fig. 3. Distribution of codon relative adaptiveness (*RA*, Eq. (1)) at different taxonomic levels.

gene sets predicted by this iterative approach with measurements of protein abundance in 29 bacterial species retrieved from PAXdb:Protein Abundance database (https://pax-db.org/) [18]. The results of this evaluation can be found in Section S2 and Table S1 of Supplementary material.

2.5. Statistical analysis

All relative adaptiveness properties examined in this study were analysed by principal component analysis (PCA) and t-distributed stochastic neighbour embedding (tSNE) using Python 2.7 with the libraries: numpy [34], pandas [35], collections (https://docs.python.org/3/library/collections.html), json (https://docs.python.org/3/library/json.html), SPARQLWrapper (https://pypi.org/project/SPARQLWrapper/), matplotlib (http://scikit-learn.sourceforge.net) [36], sklearn [37].

3. Results

3.1. Analysis of the distribution of codon relative adaptiveness index

We calculated the relative adaptiveness *RA* (Eq. (2)) for the 59 codons (excluding start and stop codons, and the tryptophan TGG codon) for each of the 4868 bacterial genomes considered using the modified OPTIMIZER algorithm. The distribution of the RA values for each AA is shown in Fig. 1; a summary of associated descriptive statistics is given in Table 1.

The broad dispersion of the *CAI* values indicates a large variability in codon bias and in *RA* values among the considered organisms. A clear bias towards one or more codons is found for some amino acids. For the amino acid aspartic acid (D) a clear preference for the codon GAT is observed, for histidine (H) is codon CAT, for isoleucine (I) is codon ATC, for asparagine (N) is codon AAC and for tyrosine (Y) is codon TAC. Some codons are hardly ever used like CAC for histidine (H), GGA for glycine (G), CTA for leucine (L) and TCG, TCA, AGT for serine (S).

To explore further the variability of codon relative adaptiveness across genomes and the relationship between different codons, we applied principal component analysis on the 59 × 4868 matrix containing the relative adaptiveness for each codon specific to each genome; a score plot of the first two principal components is given in Fig. 2. Several patterns emerge: cytosine (C) never appears in the third position of the codons located in quadrants III and IV (Fig. 2A). The corresponding amino acids are shown in Fig. 2B, although not evident patterns appear. The role of the third position is somehow expected because it is associated with the synonymous codons, and almost all synonymous codons differ at this position.

3.2. Codon usage is related to bacterial taxonomic levels

Within each bacterial taxonomic level (phylum, class, order, family and genus) we selected the three groups with the largest number of members and compared the RA values within each group as shown in Fig. 3.

In almost all cases, the median distance within the group is smaller than between groups, indicating that organisms belonging to the same group are more similar, in term of CUB, than organisms belonging to different groups; the distance between members in the same group, is smallest when the genus level (the lowest taxonomy level) is considered.

Each bacterial genome can be associated to a 59-dimensional vector of *RA* values, one for each codon. To visualize similarities between these values, t-distributed stochastic neighbour embedding (tSNE) was used to identify similarities between phyla Fig. 4A, classes Fig. 4B and order Fig. 4C. Only taxonomic levels with more than 100 representatives (genomes) were considered in this analysis.

Analysis at the phyla level (Fig. 4A) shows how Proteobacteria are scattered over the RA space; however, when clustering is performed at the class level, patterns arise (Fig. 4B): for instance, the Alpha-, Beta-, and Epsilon-proteobacteria show distinct clustering, indicating specific CUB, suggesting that each of the three Proteobacteria groups uses a similar subset of codons in their coding sequences.

Descending one taxonomic rank, to the order level, we observe distinct sub-grouping of Bacilli into Bacillales and Lactobacillales (Fig. 4C). Clustering of Burkholderiales and Campylobacterales appears to be similar to the one observed at the upper taxonomic level. We did not observe a distinct clustering for Enterobacterales which are spread all over the RA space, indicating that these bacteria adopt a large variety of codon usage.

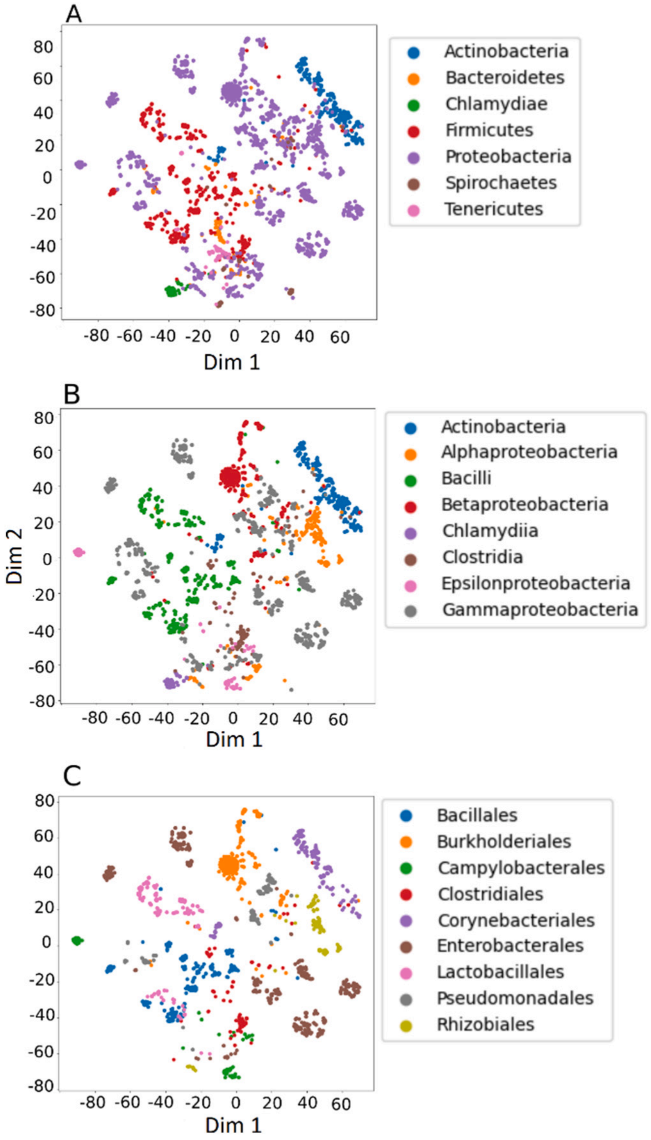


Fig. 4. The t-SNE clustering of bacterial genomes in the code RA space at different taxonomic levels: Phylum (A), class (B) and order (C). Each bacterial genome can be associated to a 59-dimensional vector of RA values, one for each codon.

3.3. Relationship between CAI, GC content and DNA polymerase III alpha subunits

Most of the bacteria from phylum Proteobacteria and all Bacteroidetes belong to dnaE1 group, most of Actinobacteria belong to dnaE1/dnaE2, and all Firmicutes belong to polC/dnaE3 group. Analysis of the association between *CAI* and genome GC content was carried separately for bacterial genomes classified according this dnaE-based grouping scheme (Table 2).

We observed that the relationship between *CAI* and GC content depends on the different usage of DNA polymerase III alpha subunits: a positive correlation between CG content and CAI is observed for dnaE1/dnaE2 genomes (Fig. 5C), while the correlation is negative for polC/dnaE3 genomes. In the case of dnaE1 genomes, the "V" shaped association between GC content and *CAI* (Fig. 5A) indicates the presence of a bimodal association, with at least two subgroups of genomes with positive and negative GC%-CAI correlation, respectively.

Bacterial genomes in the dnaE1/dnaE2 group tend to have higher average *CAI* value (0.4–0.7) and high GC content (55–75%) (Fig. 5B), while genomes belonging to the polC/dnaE3 groups (Fig. 5C) have average *CAI* values in the range of 0.3–0.7 and moderately low GC content in 30–60%. It is interesting to observe that the relationship between *CAI* and GC content is different for the different groups.

3.4. Comparison of CAI between and within protein domains

We investigated differences in CUB existing between region coding for protein functional domains (CUB in domains) and the regions coding for non-functional AA sequences. *CAI* and *RA* calculations were performed using the whole gene length for all 4868 bacterial genomes, excluding the regions just before the first domain and just after the last domain due to their negligible importance regarding codon bias calculations. Correlations between *CAI* within and between protein domains are shown in Fig. 6A: we observe that the average *CAI* in domains is higher than the average *CAI* between domains. Fig. 6B shows the same analysis at the genome level, where 7 genomes where randomly chosen (*Bacillus pseudofirmus*, *E. coli*, *Chlorobium tepidum*, *Chlamydophila pneumonium*, *Thermus thermophilus*, *Mycobacterium smegmatis*, and *Mycobacterium tuberculosis*). We observed genome-specific patterns of correlation of CAI values within/between domains. For *Mycobacterium tuberculosis* the *CAI* of codons coding for non-functional AA sequences tends to be lower than the *CAI* in domains.

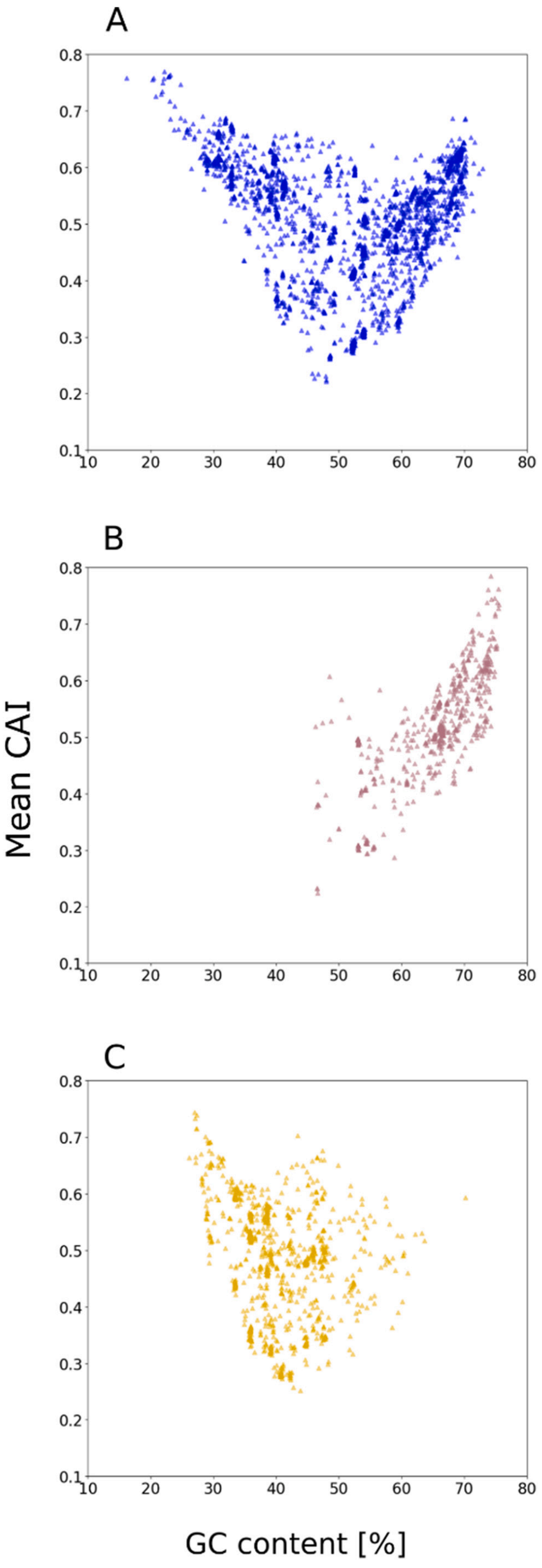
4. Discussion

We analysed codon usage across the 4868 bacterial genomes: PCA (see Fig. 1) shows how the use of (synonymous) codons is dependent on the base on the third position since almost all synonymous codons differ at this position. The third position in a codon is referred to as the "wobble position" and it plays a particular role in defining and explaining the degeneracy of the genetic code; in both prokaryotes and eukaryotes, the third position changes faster than the other two and this does not depend directly on the amino acid that a codon encodes for [38].

Cognate codon-anticodon duplexes are formed during translation through hydrogen bonding between the mRNA codon and the corresponding triplet anticodon of tRNA; between the first base of the codon

Table 2
Phyla names and numbers of bacterial genomes per each group according the dnaE scheme: dnaE1, dnaE1/dnaE2 and polC/dnaE3.

	dnaE groups					
	dnaE1	*n*	dnaE1/dnaE2	*n*	polC/dnaE3	*n*
Phyla	Actinobacteria	50	Acidobacteria	1	Firmicutes	1055
	Aquificae	5	Actinobacteria	495	Fusobacteria	8
	Bacteroidetes	139	Deinococcus-Thermus	1	Thermotogae	2
	Chlamydiae	95	Fibrobacteres	2		
	Chlorobi	9	Planctomycetes	5		
	Chloroflexi	19	Proteobacteria	411		
	Deinococcus-Thermus	13				
	Proteobacteria	2183				
	Spirochaetes	81				



(caption on next column)

Fig. 5. Correlation of the mean CAI and GC content [%]. Each point on the plot is a different bacterial genome. Genomes are classified according the different usage of polymerase III alpha sub-units and show distinct GC content patterns. A) dnaE1 group (full-spectrum), dnaE2|dnaE1 group (high-GC), and polC|dnaE3 group (low-GC).

and the third base of the anticodon, and between the second base of the codon and the second base of the anticodon, only Watson-Crick pairing (guanine–cytosine and adenine–thymine) can happen. On the contrary, there is more flexibility on the pairing between the codon third base and the anticodon first base. Crick et al. [39], suggested that more than one codon may pair with a single anticodon of the cognate transfer RNA.

The pairing at the wobble position is less precise than the pairing at the other two positions of the codon-anticodon duplex. The third base is always a major RNA base: adenine, guanine, uracil or cytosine, and each of them pairs with the first base of the anticodon which, may be guanine, uracil, cytosine or various minor RNA bases (*i.e.*, modification of regular nitrogenous bases) [40–42].

The modified nucleosides found at the wobble base in orthologous tRNA species are often different in bacteria, eukarya and archaea [41–43].

Large scale genomic analysis has led to the identification of four main decoding strategies that are diversely used in bacteria, archaea and eukarya [44]. The decoding accuracy and translation efficacy depend on specific tRNA modification enzymes that transform the canonical nucleotides of the precursor tRNA transcripts into chemically altered derivatives with innovative structural and decoding potentialities [45,46]. Certain chemical modifications are present only in specific domains or follow specific taxonomic distributions [44,47]. Reconstruction of ancestral tRNA modifications across bacterial genomes has shown that most modifications were ancestral to eubacteria but lost in many lineages; those losses coincided with evolutionary shifts in non-target tRNA and were found to be driven by bias in genomic GC content and associated codon use [48].

Large scale analyses, like the one here presented, are empowered by the increasing availability of high-quality data and bioinformatic tools solely relying on sequence data. Here, the OPTIMIZER algorithm allowed the use of *RA* and *CAI* to evaluate CUB.

In the same way, the tool named stAI$_{calc}$ [49,50] can be used to compute the tRNA adaptation index (*tAI*) [16] solely based on sequence information. The *tAI* provides an alternative measurement of CUB that explicitly incorporates the dynamics of tRNA and mRNA binding and could be especially relevant when complemented with tRNA detection tools that would further shed light on the association between tRNA, tRNA copy number and CUB. However, this link has been reported to be weaker in slower growing bacterial species and the predictive value of tRNAs and their copy number to predict CUB in these species appears to be limited [51–53] unless complemented with expression data [53] which prevents the application of this index at large scale.

While codon usage can differ not only between organisms but also within different regions of a genome and even within a gene [54,55], it is known that a specific codon usage characterizes each bacterial species and that the majority of its genes shares such bias [55,56], implying that the genome, not the individual gene, is the unit of selection. Moreover, it has been suggested that codon bias in specialized categories of genes is a re-modulation of the distinctive codon bias of the species [57].

Bacterial genes undergo translational selection, and highly expressed genes use codons that are translated faster and/or more accurately by the ribosome [2,56]. The relative adaptation index captures this bias towards translational efficiency, and our results show that this is strongly associated with bacterial taxonomy (Figs. 4 and 5).

Association between CUB and bacterial adaptation has been suggested [58,59], and it has been shown that species with given phenotypic traits and living in similar environmental conditions have similar codon preferences [57]: this suggests an evolutionary convergence of

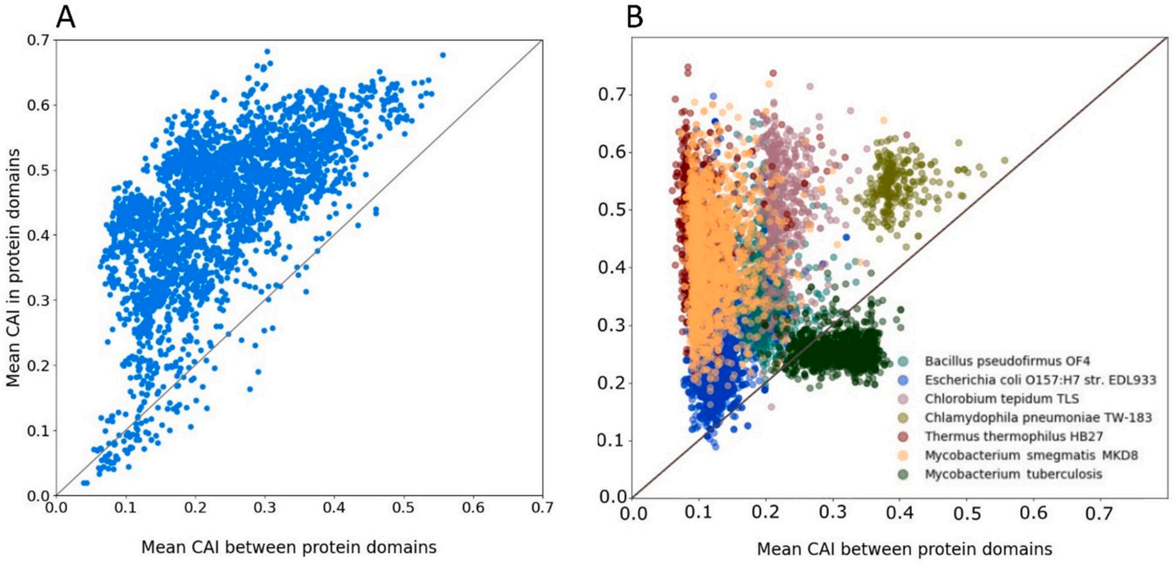


Fig. 6. A) Correlation between mean CAI in gene regions coding for protein functional domains and mean CAI in region between domains for the 4868 bacterial genomes. Average is taken at the genome level. Each point on the graph represents one of 4868 bacterial genomes selected. B) Correlation between mean CAI in gene regions coding for protein functional domains and mean CAI in region between domains for 7 randomly selected bacterial genomes.

CUB and adaptation in groups of organisms sharing similar physiology and/or living in similar habitats.

We observed Proteobacteria having very diverse CUB and *RA* profiles (Fig. 4), with the existence of many bacterial groups sharing very similar codon bias and RA characteristics. The existence of sub-groups within the six taxonomic classes reflects the great phenotypic variability observed in this phylum that includes, among others, pathogenic (like *Escherichia, Salmonella, Vibrio*) and free-living nitrogen-fixing bacteria [60]. This is also consistent with the observation that phenotypic traits, rather than phylogenetic relatedness, underlie the similarities in CUB between organisms [59].

Several processes have been suggested to contribute to the codon adaptation, like lateral gene transfer mechanisms which drive convergence to environmental parameters (like pressure, salinity and temperature) or driving adaptation to the abundance or lack of nutrients, or, in pathogens, driving the adaptation to the host through mechanisms developed to escape the host immune response [60].

In our analysis we observed a strong relationship between *CAI* and bacterial GC content (Fig. 5). The GC content of bacterial genomes ranges from 25% to 75% [61,62] and already in the 1960's it was hypothesized that the GC content could affect the protein AA sequence [62,63]. Several studies have shown that genomic GC content is correlated with gene single base frequencies [64,38] as well with the frequency of amino acids [62,65,66].

Codon usage and protein amino acid content have evolved independently in different groups of bacteria. The use of amino acids encoded by GC-rich codons increases by approximately 1% for each 10% increase in genomic GC content, suggesting that GC content is the primary determinant of the association between AA usage and codons. The AA usage patterns observed in bacterial genomes and the selection for translational efficiency of highly expressed genes are constrained by the genomic determinants associated with the GC content [66]. Similar base usage patterns at the three codon positions, codon usage patterns, and amino acid usage patterns have been observed in distant phylogenetic lineages sharing similar GC content, indicating that GC content results in similar codon bias regardless of phylogenetic lineages [67].

We found the association between CUB and GC content to be dependent on which dimeric combination of DNA polymerase III alpha sub-units the bacteria adopt. Bacterial genomes are classified into three groups with distinct GC content variation spectra: dnaE1 (full-spectrum), dnaE2|dnaE1 (high-GC), and polC|dnaE3 (low-GC) [18], and it has been suggested that DNA polymerase III alpha sub-units and their isoform may play a pivotal role in determining GC variability, while environmental or bacteriological factors, such as genome size, temperature, oxygen requirement, and habitat, either play subsidiary roles or rely indirectly on different mutator genes to fine-tune the GC content [14].

Our results are fully consistent with this observation: distantly related bacteria belong to the same dnaE-based grouping, indicating similar GC variability, which in turn drives similar codon bias. We observed a positive correlation between codon adaptation and GC content in dnaE1|dnaE2 bacteria (including among other Actinobacteria and Proteobacteria), while negative correlation is observed for polC|dnaE bacteria (including Firmicutes, Fusobacteria and Thermotogae). This is consistent with the fact that phenotypic traits rather than phylogenetic relatedness underlie the similarities in the codon usage [14,68]. For instance, aerobe bacteria carry the GC-enriching polymerase, while anaerobes carry the AT-enriching polymerases, an example of which can be found, for instance, among Proteobacteria. Moreover, within the three groups (dnaE1 (full-spectrum), dnaE2|dnaE1 (high-GC), and polC|dnaE3 (low-GC)), the GC content is found to correlate with optimal growth temperature [14], a phenotypic trait associated with differential codon usage [14,68,69]. Most terrestrial, plant-associated, and nitrogen-fixing bacteria are from dnaE1|dnaE2 group. In contrast, most pathogenic or symbiotic bacteria in insects, and those living in aquatic environments, belong to the dnaE1|polV group [14].

We showed that preferred codons are mainly in the domains (Fig. 6) and *CAI* values in the domains are significantly higher than between domains which is in line with the hypothesis that codon usage have an impact on translation speed [19,70], with translation faster in domains, because of the great number of preferred codons [71].

It is not clear, at this stage, if such large differences of CAI values in within/between proteins domain could be accounted by normalizing for the amount of disorder in the genome or in the protein sequence, considering, for instance, the ratio between the length of region predicted to be unstructured to the total gene or protein length. From a structural point of view, it has been shown that in eukaryotes [18] and fungi [72], the preferred codons are mostly used to encode α-helices and β-sheets regions while non-optimal codons codes for coiled-coli regions.

Data availability

All data were obtained from the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home). Supplementary material and code for analysis can be found at https://gitlab.com/wurssb/cai.

Funding

This work was partially supported by the Dutch National Roadmap Bridge funding, Project number NRGWI.obrug.2018.005 and by The Netherlands Organisation for Health Research and Development (ZonMW) through the PERMIT project (Personalized Medicine in Infections: from Systems Biomedicine and Immunometabolism to Precision Diagnosis and Stratification Permitting Individualized Therapies, project number 456008002) under the PerMed Joint Transnational call JTC 2018 (Research projects on personalised medicine—smart combination of pre-clinical and clinical research with data and ICT solutions). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Credit authorship contribution statement

M.S.-D. and E.S. conceived the study and provided supervision. A.M.-G. and M.R.M. vd B. performed analysis. A.M.-G. drafted the manuscript. M.S.-D. and E.S. revised the manuscript in its final form. All authors approved the manuscript.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbagrm.2022.194826.

References

[1] U. Lagerkvist, “Two out of three”: an alternative method for codon reading, Proc. Natl. Acad. Sci. 75 (1978) 1759–1762.
[2] M. Gouy, C. Gautier, Codon usage in bacteria: correlation with gene expressivity, Nucleic Acids Res. 10 (1982) 7055–7074.
[3] R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pave, Codon catalog usage and the genome hypothesis, Nucleic Acids Res. 8 (1980) 197.
[4] D.A. Drummond, C.O. Wilke, Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution, Cell 134 (2008) 341–352.
[5] O. Man, J.L. Sussman, Y. Pilpel, Examination of the tRNA adaptation index as a predictor of protein expression levels, in: Systems Biology and Regulatory Genomics, Springer, 2005, pp. 107–118.
[6] S. Andersson, C. Kurland, Codon preferences in free-living microorganisms, Microbiol. Rev. 54 (1990) 198–210.
[7] T. Ikemura, H. Ozeki, Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents, in: Cold Spring Harbor Symposia on Quantitative Biology vol. 47, Cold Spring Harbor Laboratory Press, 1983, pp. 1087–1097.
[8] T. Ikemura, Codon usage and tRNA content in unicellular and multicellular organisms, Mol. Biol. Evol. 2 (1985) 13–34.
[9] P.M. Sharp, W.-H. Li, The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications, Nucleic Acids Res. 15 (1987) 1281–1295.
[10] P.M. Sharp, W.-H. Li, Codon usage in regulatory genes in Escherichia coli does not reflect selection for ‘rare’ codons, Nucleic Acids Res. 14 (1986) 7737–7749.
[11] A. Carbone, A. Zinovyev, F. Képes, Codon adaptation index as a measure of dominating codon bias, Bioinformatics 19 (2003) 2005–2015.
[12] Y. Harigaya, R. Parker, The link between adjacent codon pairs and mRNA stability, BMC Genomics 18 (2017) 1–16.
[13] L.A. Diambra, Differential bicodon usage in lowly and highly abundant proteins, PeerJ 5 (2017), e3081.
[14] H. Wu, Z. Zhang, S. Hu, J. Yu, On the molecular mechanism of GC content variation among eubacterial genomes, Biol. Direct 7 (2012) 1–16.
[15] X. Zhao, Z. Zhang, J. Yan, J. Yu, GC content variability of eubacteria is governed by the pol III α subunit, Biochem. Biophys. Res. Commun. 356 (2007) 20–25.
[16] M.d. Reis, R. Savva, L. Wernisch, Solving the riddle of codon usage preferences: a test for translational selection, Nucleic Acids Res. 32 (2004) 5036–5044.
[17] K. Homma, T. Noguchi, S. Fukuchi, Codon usage is less optimized in eukaryotic gene segments encoding intrinsically disordered regions than in those encoding structural domains, Nucleic Acids Res. 44 (2016) 10051–10061.
[18] M. Zhou, T. Wang, J. Fu, G. Xiao, Y. Liu, Nonoptimal codon usage influences protein structure in intrinsically disordered regions, Mol. Microbiol. 97 (2015) 974–987.
[19] R. Saunders, C.M. Deane, Synonymous codon usage influences the local protein structure observed, Nucleic Acids Res. 38 (2010) 6719–6728.
[20] J.J. Koehorst, J.C. van Dam, E. Saccenti, V.A. Martins dos Santos, M. Suarez-Diez, P.J. Schaap, SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles, Bioinformatics 34 (2018) 1401–1403.
[21] D. Hyatt, G.-L. Chen, P.F. LoCascio, M.L. Land, F.W. Larimer, L.J. Hauser, Prodigal: prokaryotic gene recognition and translation initiation site identification, BMC Bioinforma. 11 (2010) 1–11.
[22] D.H. Haft, J.D. Selengut, R.A. Richter, D. Harkins, M.K. Basu, E. Beck, TIGRFAMs and genome properties in 2013, Nucleic Acids Res. 41 (2012) D387–D395.
[23] A.N. Nikolskaya, C.N. Arighi, H. Huang, W.C. Barker, C.H. Wu, PIRSF family classification system for protein functional and evolutionary analysis, Evol. Bioinforma. 2 (2006), 117693430600200033.
[24] J. Schug, S. Diskin, J. Mazzarelli, B.P. Brunk, C.J. Stoeckert, Predicting gene ontology functions from ProDom and CDD protein domains, Genome Res. 12 (2002) 648–655.
[25] I. Letunic, R.R. Copley, B. Pils, S. Pinkert, J. Schultz, P. Bork, SMART 5: domains in the context of genomes and networks, Nucleic Acids Res. 34 (2006) D257–D260.
[26] C.J. Sigrist, L. Cerutti, E. De Castro, P.S. Langendijk-Genevaux, V. Bulliard, A. Bairoch, N. Hulo, PROSITE, a protein domain database for functional characterization and annotation, Nucleic Acids Res. 38 (2010) D161–D166.
[27] J. Bolleman, E. de Castro, D. Baratin, S. Gehant, B.A. Cuche, A.H. Auchincloss, E. Coudert, C. Hulo, P. Masson, I. Pedruzzi, HAMAP as SPARQL rules—a portable annotation pipeline for genomes and proteomes, GigaScience 9 (2020), giaa003.
[28] A. Bateman, L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L. Sonnhammer, The Pfam protein families database, Nucleic Acids Res. 32 (2004) D138–D141.
[29] T.K. Attwood, M.J. Blythe, D.R. Flower, A. Gaulton, J. Mabey, N. Maudling, L. McGregor, A.L. Mitchell, G. Moulton, K. Paine, PRINTS and PRINTS-S shed light on protein ancestry, Nucleic Acids Res. 30 (2002) 239–241.
[30] D. Wilson, M. Madera, C. Vogel, C. Chothia, J. Gough, The SUPERFAMILY database in 2007: families and functions, Nucleic Acids Res. 35 (2007) D308–D313.
[31] C. Yeats, J. Lees, A. Reid, P. Kellam, N. Martin, X. Liu, C. Orengo, Gene3D: comprehensive structural and functional annotation of genomes, Nucleic Acids Res. 36 (2007) D414–D418.
[32] J.C. van Dam, J.J. Koehorst, J.O. Vik, V.A. Martins dos Santos, P.J. Schaap, M. Suarez-Diez, The Empusa code generator and its application to GBOL, an extendable ontology for genome annotation, Sci. Data 6 (2019) 1–9.
[33] P. Puigbo, E. Guzman, A. Romeu, S. Garcia-Vallve, OPTIMIZER: a web server for optimizing the codon usage of DNA sequences, Nucleic Acids Res. 35 (2007) W126–W131.
[34] C.R. Harris, K.J. Millman, S.J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, Array programming with NumPy, Nature 585 (2020) 357–362.
[35] W. McKinney, Data structures for statistical computing in python, in: Proceedings of the 9th Python in Science Conference vol. 445, 2010, pp. 51–56. Austin, TX.
[36] J.D. Hunter, Matplotlib: a 2D graphics environment, Comput. Sci. Eng. 9 (2007) 90–95.
[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[38] R.D. Knight, S.J. Freeland, L.F. Landweber, A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes, Genome Biol. 2 (2001) 1–13.
[39] F. Crick, Codon—Anticodon Pairing: The Wobble Hypothesis, 1966.
[40] G. Das, R.H.D. Lyngdoh, Role of wobble base pair geometry for codon degeneracy: purine-type bases at the anticodon wobble position, J. Mol. Model. 18 (2012) 3805–3820.
[41] H. Grosjean, M. Sprinzl, S. Steinberg, Posttranscriptionally modified nucleosides in transfer RNA: their locations and frequencies, Biochimie 77 (1995) 139–141.
[42] H. Grosjean, Nucleic acids are not boring long polymers of only four types of nucleotides: a guided tour, in: Madame Curie Bioscience Database, Landes Bioscience, 2013 (Internet).
[43] G.R. Björk, J.U. Ericson, C.E. Gustafsson, T.G. Hagervall, Y.H. Jönsson, P. M. Wikström, Transfer RNA modification, Annu. Rev. Biochem. 56 (1987) 263–285.
[44] H. Grosjean, V. de Crécy-Lagard, C. Marck, Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes, FEBS Lett. 584 (2010) 252–264.
[45] O. Namy, F. Lecointe, H. Grosjean, J.-P. Rousset, Translational recoding and RNA modifications, in: Fine-tuning of RNA Functions by Modification and Editing, Springer, 2005, pp. 309–340.
[46] P.F. Agris, F.A. Vendeix, W.D. Graham, tRNA’s wobble decoding of the genome: 40 years of modification, J. Mol. Biol. 366 (2007) 1–13.
[47] M.A. Machnicka, K. Milanowska, O. Osman Oglou, E. Purta, M. Kurkowska, A. Olchowik, W. Januszewski, S. Kalinowski, S. Dunin-Horkawicz, K.M. Rother, MODOMICS: a database of RNA modification pathways—2013 update, Nucleic Acids Res. 41 (2012) D262–D267.

[48] G.D. Diwan, D. Agashe, Wobbling forth and drifting back: the evolutionary history and impact of bacterial tRNA modifications, Mol. Biol. Evol. 35 (2018) 2046–2059.

[49] R. Sabi, T. Tuller, Modelling the efficiency of codon–tRNA interactions based on codon usage bias, DNA Res. 21 (2014) 511–526.

[50] R. Sabi, R. Volvovitch Daniel, T. Tuller, stAIcalc: tRNA adaptation index calculator based on species-specific weights, Bioinformatics 33 (2017) 589–591.

[51] J. Rojas, G. Castillo, L.E. Leiva, S. Elgamal, O. Orellana, M. Ibba, A. Katz, Codon usage revisited: lack of correlation between codon usage and the number of tRNA genes in enterobacteria, Biochem. Biophys. Res. Commun. 502 (2018) 450–455.

[52] E.P. Rocha, Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization, Genome Res. 14 (2004) 2279–2286.

[53] Y. Wei, J.R. Silke, X. Xia, An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria, Sci. Rep. 9 (2019) 1–11.

[54] S.D. Hooper, O.G. Berg, Gradients in nucleotide and codon usage along Escherichia coli genes, Nucleic Acids Res. 28 (2000) 3517–3523.

[55] J.B. Plotkin, G. Kudla, Synonymous but not the same: the causes and consequences of codon bias, Nat. Rev. Genet. 12 (2011) 32–42.

[56] J.L. Bennetzen, B.D. Hall, Codon selection in yeast, J. Biol. Chem. 257 (1982) 3026–3031.

[57] M. Dilucca, G. Cimini, A. Semmoloni, A. Deiana, A. Giansanti, Codon bias patterns of E. coli's interacting proteins, PLoS One 10 (2015), e0142127.

[58] H. Jiang, W. Guan, D. Pinney, W. Wang, Z. Gu, Relaxation of yeast mitochondrial functions after whole-genome duplication, Genome Res. 18 (2008) 1466–1471.

[59] M. Botzman, H. Margalit, Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles, Genome Biol. 12 (2011) 1–11.

[60] A. Carbone, F. Kepes, A. Zinovyev, Codon bias signatures, organization of microorganisms in codon space, and lifestyle, Mol. Biol. Evol. 22 (2005) 547–561.

[61] A.N. Belozersky, A.S. Spirin, A correlation between the compositions of deoxyribonucleic and ribonucleic acids, Nature 182 (1958) 111–112.

[62] N. Sueoka, Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein, Proc. Natl. Acad. Sci. U. S. A. 47 (1961) 1141.

[63] N. Sueoka, On the genetic basis of variation and heterogeneity of DNA base composition, Proc. Natl. Acad. Sci. U. S. A. 48 (1962) 582.

[64] A. Muto, S. Osawa, The guanine and cytosine content of genomic DNA and bacterial evolution, Proc. Natl. Acad. Sci. 84 (1987) 166–169.

[65] J. Lobry, Influence of genomic G+ C content on average amino-acid composition of proteins from 59 bacterial species, Gene 205 (1997) 309–316.

[66] J. Lightfield, N.R. Fram, B. Ely, Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage, PLoS One 6 (2011), e17677.

[67] H.-Q. Zhou, L.-W. Ning, H.-X. Zhang, F.-B. Guo, Analysis of the relationship between genomic GC content and patterns of base usage, codon usage and amino acid usage in prokaryotes: similar GC content adopts similar compositional frequencies regardless of the phylogenetic lineages, PLoS One 9 (2014), e107319.

[68] D. Arella, M. Dilucca, A. Giansanti, Codon usage bias and environmental adaptation in microbial organisms, Mol. Gen. Genomics. (2021) 1–12.

[69] J. Lobry, A. Necşulea, Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes, Gene 385 (2006) 128–136.

[70] G. Hanson, J. Coller, Codon optimality, bias and usage in translation and mRNA decay, Nat. Rev. Mol. Cell Biol. 19 (2018) 20–30.

[71] C.-H. Yu, Y. Dang, Z. Zhou, C. Wu, F. Zhao, M.S. Sachs, Y. Liu, Codon usage influences the local rate of translation elongation to regulate co-translational protein folding, Mol. Cell 59 (2015) 744–754.

[72] S. Pechmann, J. Frydman, Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding, Nat. Struct. Mol. Biol. 20 (2013) 237–243.