Codon frequency is modulated by proteic selection, resulting in a coding profile in Archaea and Yeast

Abstract

Codons as fragments of the genetic code articulate both nucleotidic and proteic constraints. If codon usage bias is now admitted to be mainly influenced by GC content, codon frequencies in general may display a more subtle compromise between base composition and selection at proteic level. In order to investigate the existing non-GC content factors of codon frequencies, we compared coding sequences (CDS) of 280 Archaea plus *S. cerevisiae* genomes to their randomized version (same base-composition and same length). Through dedicated counts we identified several CDS vs random patterns in Archaea some of which reflecting probable or evident proteic constraint : in particular, the systematic enrichment of CDS in negatively charged amino acids, and the strong constraint existing on codons having a T in second position, which, on the basis of hydrophobic cluster analysis attests a folding constraint. The sum of these patterns constitutes a coding profile that enables to accurately classify about 99% of individual archaea sequences between CDS and randomized CDS. In *S. cerevisiae*, whose coding profile shares similarities with Archeae of close GC content, phylostratigraphic methods allowed to investigate the coding profile of CDS based on their relative age. This analysis reveals that contrary to other genes, the youngest genes (only found in *S. cerevisiae*) as a whole do not have a strong coding profile. This can be explained by their relative shortness in comparison with other genes. But even when taking length into account, a clear enrichment of misclassified sequences appears in the youngest *S. cerevisiae* genes. This enrichment may reflect an insufficient proteic optimization operated by selection.

Introduction

The very existence of a genetic code implies, as any code, the transmission of information between at least two parties. Thus, genetic code lies between two biological objects : the genome, and the proteome. Codons as the interpretation units of genetic code are largely studied from a codon usage bias point of view. Indeed, all amino acids but tryptophan and methionine are encoded by more than one codon [1]. This genetic code degeneracy allows the existence of synonymous mutations and so synonymous codons. When counting these synonymous codons in a given genome, genomic region, or gene, it appears a codon usage bias e.i an unbalanced count between some codons and their synonyms.

Codon bias studies revealed its correlation with tRNA abundance and translational efficiency, offering many valuable perspectives [2]. Although it has been suggested that codon bias was mainly influenced by tRNA abundance, it is now clear that codon bias is mostly shaped by GC content [3]. Genome wide GC content is indeed highly correlated with codon bias and necessarily causative of it since genome-wide GC content is also highly correlated with non-coding genome GC content. Thus, under a single codon bias perspective we observe a main GC content cause with primarily translational consequences.

Yet, if we aim to fully acquire knowledge about codon usage, we need to consider it as it is : a conceptual articulation between a triplet of nucleotides and an amino-acid (through tRNA). As a direct consequence the frequency of the 64 codons is not imposed by base composition (which according to the second Chargaff's parity rule can be fairly represented by GC content among double stranded DNA genomes [4]). Codon frequencies, as we observe them, are the result of both nucleotide and protein constraints. Even models that aim to predict codon bias (only between synonymous codons) from genome GC content show limitations, especially in certain phyla [3]. But when considering the entire codon table it clearly appears that some codons or groups of codons undergo strong non-GC constraints (as an illustration see in the Results section : figure 7).

Now, considering that the codon frequencies of coding sequences (CDS) result from :

$$CF = GCconstraint + nonGCconstraints$$

then

$$nonGCconstraints = CF - GCconstraint$$

with CF corresponding to Codon Frequencies. Thus to isolate non-GC constraints one needs to access codon frequencies resulting from purely GC constraints. For this study we used 280 Archaea genomes which provided CDS bearing both GC and non-GC constraints. So we needed to obtain sequences only built on the GC constraint, in order to compare them with CDS and by contrast, to consider nonGC constraints affecting the CDS.

1st base	**2nd base**								**3rd base**
	T		**C**		**A**		**G**		
T	TTT	Phe	CTT	Ser	ATT	Tyr	GTT	Cys	**T**
	TTC		CTC		ATC		GTC		**C**
	TTA	Leu	CTA		ATA	Stop (Ochre)	GTA	Stop (Opale)	**A**
	TTG		CTG		ATG	Stop (Ambre	GTG	Trp	**G**
C	TCT		CCT	Pro	ACT	His	GCT	Arg	**T**
	TCC		CCC		ACC		GCC		**C**
	TCA		CCA		ACA	Gln	GCA		**A**
	TCG		CCG		ACG		GCG		**G**
A	TAT	Ile	CAT	Thr	AAT	Asn	GAT	Ser	**T**
	TAC		CAC		AAC		GAC		**C**
	TAA		CAA		AAA	Lys	GAA	Arg	**A**
	TAG	Met	CAG		AAG		GAG		**G**
G	TGT	Val	CGT	Ala	AGT	Asp	GGT	Gly	**T**
	TGC		CGC		AGC		GGC		**C**
	TGA		CGA		AGA	Glu	GGA		**A**
	TGG		CGG		AGG		GGG		**G**

table 1 : Standard DNA codon table, adapted from https://en.wikipedia.org/wiki/DNA_and_RNA_codon_tables

This is why in this study we compared the DNA coding sequences (CDS) of 280 archeal genomes with their randomized nucleotidic sequences, which conserved the exact same base composition and length as the original sequences.

Material and methods

Access to sequences

The fasta files containing the original CDS of the 280 archaeal genomes as well as the *S. cerevisiae* genome used for this study were generated with ORFmine [5]. "*ORFmine is an open-source package that aims at extracting, annotating, and characterizing the fold potential and the structural properties of all Open Reading Frames (ORF) encoded in a genome (including coding and noncoding sequences).*"

The fasta file containing the original CDS sequences and the fasta file containing the randomized CDS are available for *S. cerevisiae* and three example Archaea genomes at : https://drive.google.com/drive/folders/1h5iwwB_v4C-rYl_7kd-OBXJaNogrvDQP?usp=sharing.
Codon counts were performed with the help of the coRdon R package [6].

Sequence classification

Sequence classifications were done using the XGBoost implementation of Scikit-learn. The choice of this algorithm is based on both its predicting performance, its convenience and its speed in reproducing the result.
Features used were the mean base content, the mean relative frequency of each codon, amino acids, and same-base, same-position, and XYN groups of codons as described in the Results section.
Stop codon frequencies were removed from codon count prior to any other computation, since a typical CDS sequence should always possess only one stop codon.

All available sequences were used for each model, which implies half CDS and half randomized CDS. Apart from 10 K-fold cross validations, 80% of data was used for training and 20% for testing.

Although it was possible to slightly improve them with some hyperparameters tuning, the current results can be reproduced with a default setting.
Accuracies followed by a standard deviation (sd) value refer to 10-fold cross-validation results.

Finally, the count of misclassified sequences for *S. cerevisiae* was done by gathering the tested sequences of all 10-fold cross-validation.

Genomic phylostratigraphy

Genomic phylostratigraphy is based both on a phylogenetic tree and an homologous sequence detection method. The relative distance of the species on the tree to the species of interest is used to assign each homologous sequence its relative "age" (or "genomic phylostrata"). Here the dated genes used come from Wilson et al (2017) [7], in which the authors applied a BLASTp search on Saccharomyces Genome Database (E-value threshold : 0.001 NCBI non-redundant protein sequences (nr) database.

Phylostratum	Gene shared among	Number of genes
10	*Saccharomyces cerevisiae*	1092
9	Saccharomycetaceae	347
8	Saccharomycetales	366
7	Saccharomyceta	71
6	Ascomycota	140
5	Dikarya	77
4	Fungi	290
3	Opisthokonta	127
2	Eukaryota	1579
1	Cellular_Organisms	2592
0	No assignation	11

table 2 : A key to the phylostratum numbers used for CDS datation.

Others

The fisher exact test p-value was computed with the stats R package [8].

All R and python scripts used written for this study are available at : https://github.com/SnoopBZH/FMP.

Results

Systematic CDS deviation from random at AA and codon table levels

Although correlated with GC content, amino acid frequencies in CDS can not be fully explained by a simple base frequency input. Figure 1 illustrates how the amino acid distribution deviates from what is expected by chance. Furthermore, the randomized CDS IQR wideness highlights how much amino acids are sensitive to GC content variation.

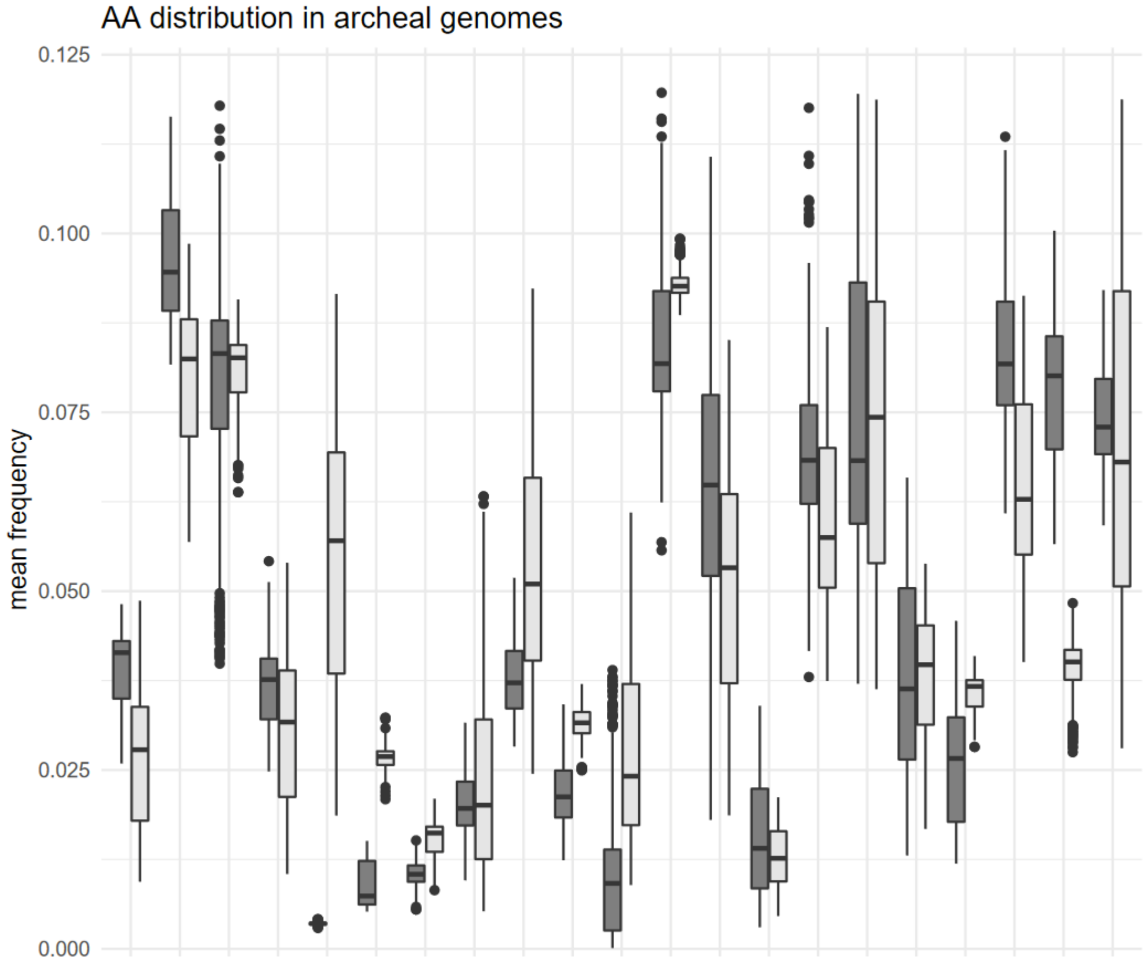
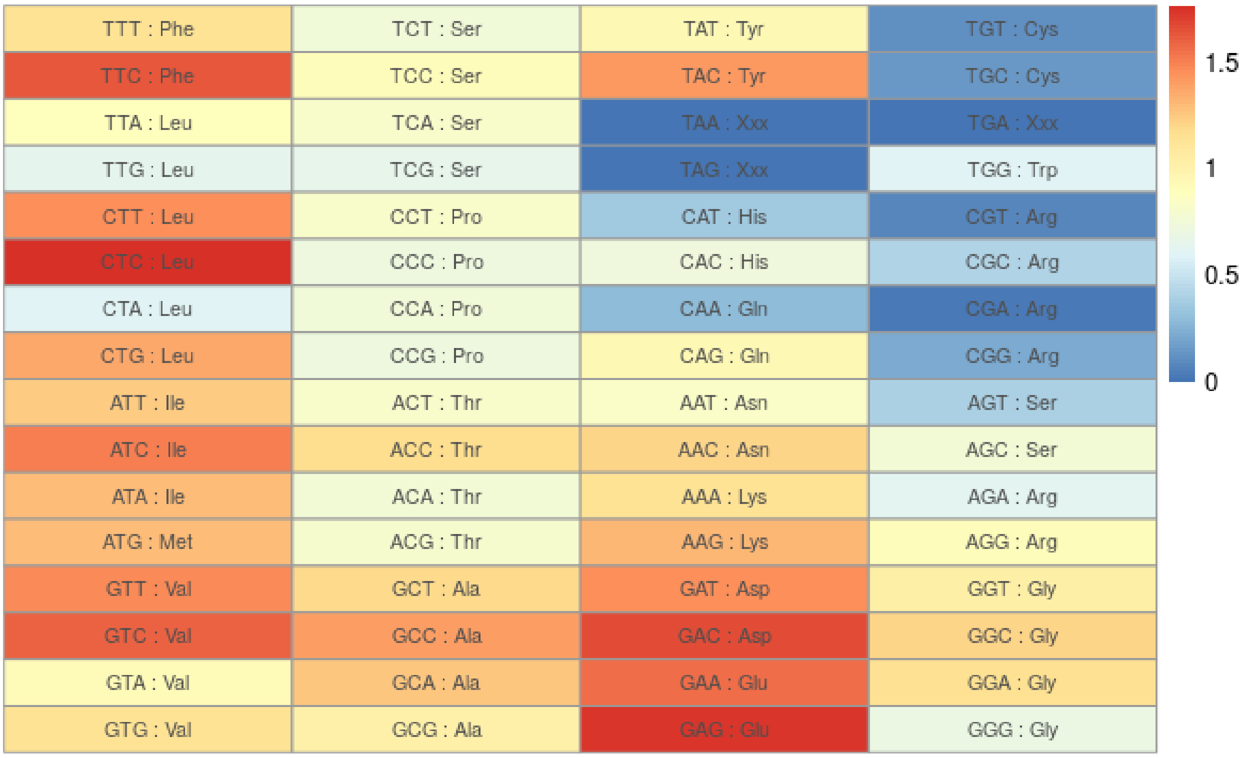


figure 1 : Fore each amino acid, the distribution of its mean frequency (one value per genome) among the CDS of 280 Archeae genomes (left boxplot) and their randomized version (right boxplot)



TTT : Phe	TCT : Ser	TAT : Tyr	TGT : Cys
TTC : Phe	TCC : Ser	TAC : Tyr	TGC : Cys
TTA : Leu	TCA : Ser	TAA : Xxx	TGA : Xxx
TTG : Leu	TCG : Ser	TAG : Xxx	TGG : Trp
CTT : Leu	CCT : Pro	CAT : His	CGT : Arg
CTC : Leu	CCC : Pro	CAC : His	CGC : Arg
CTA : Leu	CCA : Pro	CAA : Gln	CGA : Arg
CTG : Leu	CCG : Pro	CAG : Gln	CGG : Arg
ATT : Ile	ACT : Thr	AAT : Asn	AGT : Ser
ATC : Ile	ACC : Thr	AAC : Asn	AGC : Ser
ATA : Ile	ACA : Thr	AAA : Lys	AGA : Arg
ATG : Met	ACG : Thr	AAG : Lys	AGG : Arg
GTT : Val	GCT : Ala	GAT : Asp	GGT : Gly
GTC : Val	GCC : Ala	GAC : Asp	GGC : Gly
GTA : Val	GCA : Ala	GAA : Glu	GGA : Gly
GTG : Val	GCG : Ala	GAG : Glu	GGG : Gly

figure 2 : Difference of mean frequency between CDS and randomized CDS. Values were transformed into an exponential scale.

Figure 2 represents for each codon its difference of mean frequency between CDS and randomized CDS (represented in an exponential scale). Prior to other considerations this representation points out that the codon table (and so the genetic code) is far to be randomly organized, especially in that ensuring the repartition of the first two codon bases allows a lot of base variability at the third position. The present tendencies are diluted in a wide range of GC contents. Indeed, even if the difference between a CDS and its randomized version attests of the nonGC component of codon frequencies, these nonGC constraints may vary according to the GC content, for example to maintain a given amino acid at a certain level. Each of the 64 codons genome mean frequency correlation between CDS and randomized CDS can be found in the Supplementary Material section. It confirms that as expected, many codons can be depleted in CDS at a given GC content level while enriched at another level.

CDS vs random patterns trough codons grouping

The codon heatmap hotspots (Figure 2) point out the opportunity of studying the differential usage of several groups of codons among CDS and randomized CDS.

In this sense, in addition with AA frequency and codon frequency we performed three additional counts based on two different codon grouping :

First, when summing all the codons sharing a given base at a given triplet position, we performed two counts : the same-position count is the relative frequency of each of the four bases (T/C/A/G) at a given codon position. For example “1T = 0.15” means that 15% of codons start with a T. Then the same-base count is the relative frequency of a given base across the three positions. “T1 = 0.25” means that 25% of T bases are found at the first position of the codons. Then, the XYN count refers to the sum of codons sharing the same first two bases (e.g. GAT, GAC, GAA and GAG.

First of all, the same-base count reveals how for each of the four possible bases, the third position constitutes an adjustment variable in comparison with the first two positions which are much more constrained since they mainly bear the amino acid property (see table 1 and figure 4).

Then, 1T comparison between CDS and randomized CDS shows a systematic depletion in CDS that grows with AT content. This signal is mainly due to the presence in the TNN column of six codons :

- The three stop codons because of the necessity to have a minimum number of amino acid residues encoded in each CDS, the three stop codons are dramatically less frequent than they should be by chance (e.i. in their randomized version). As an illustration, in a genome with 50% GC content, generating by chance a sequence of 100 residues before a stop is (from the binomial law)

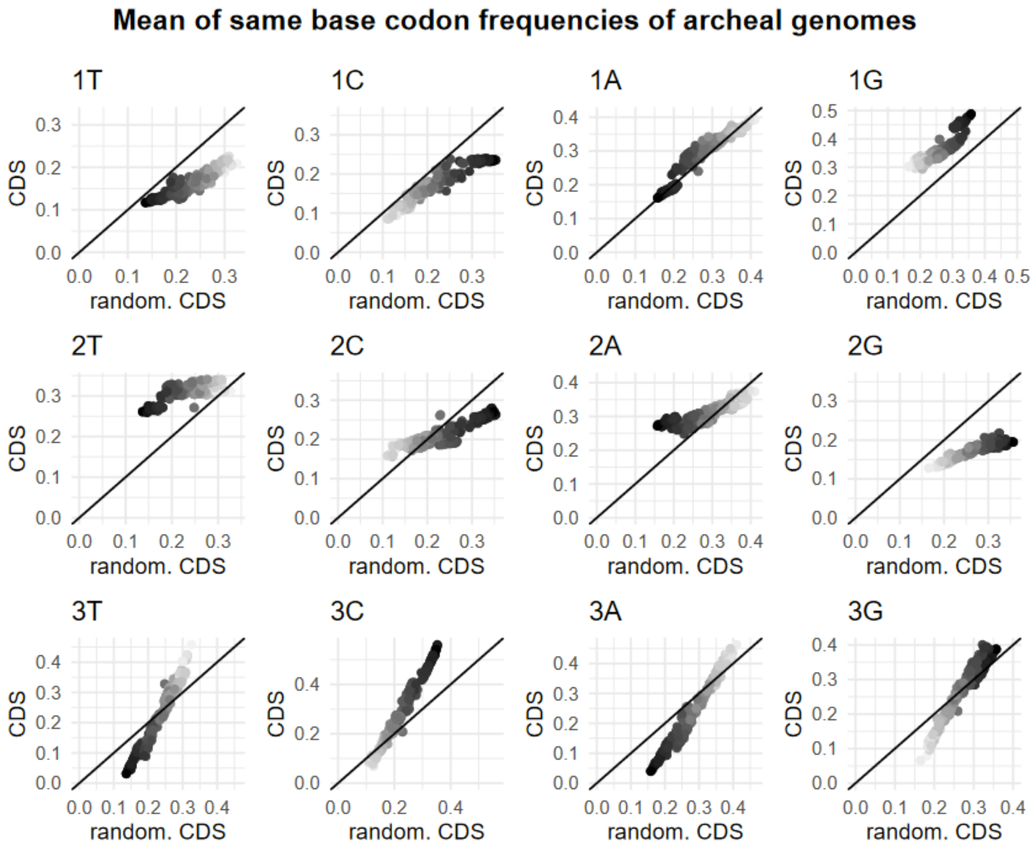


figure 3 : Correlation between the same-position count in CDS vs randomized CDS.

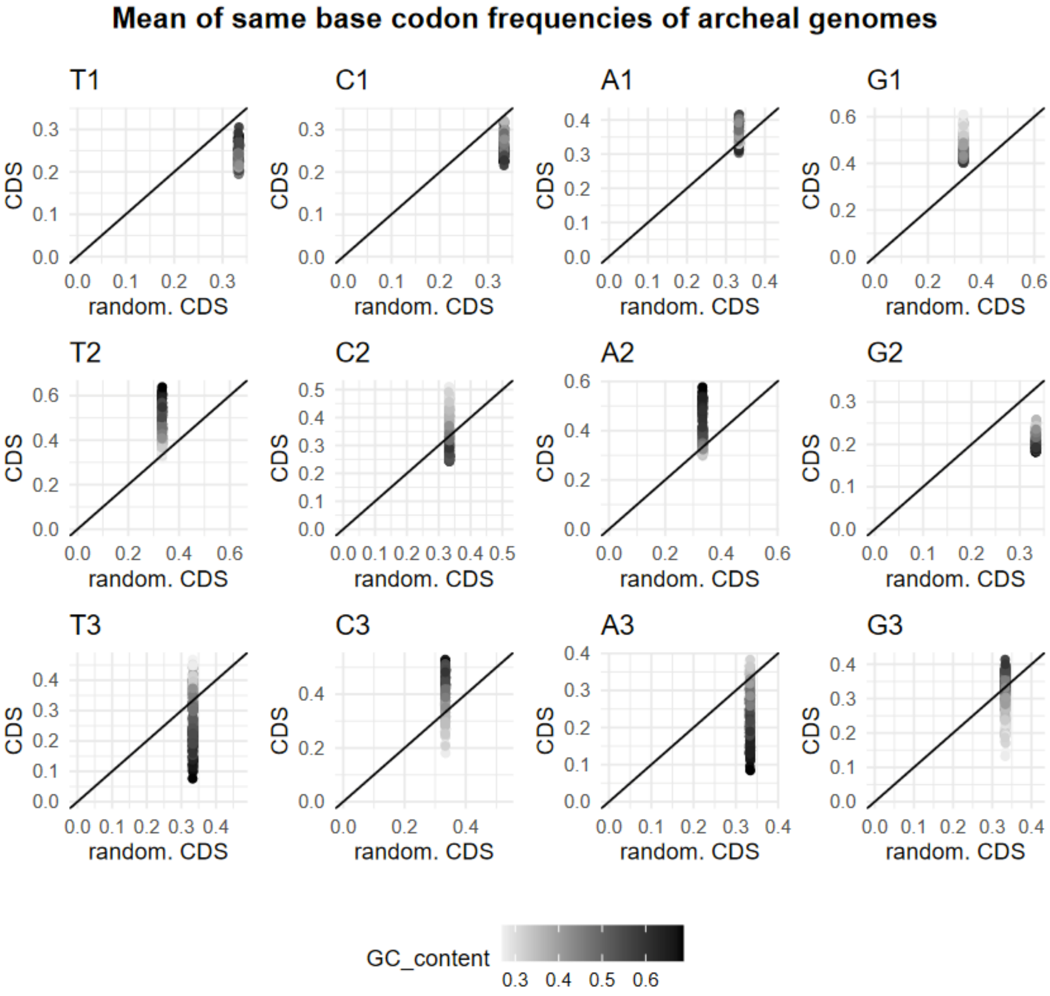


figure 4 : Correlation between the same-position count in CDS vs randomized CDS.

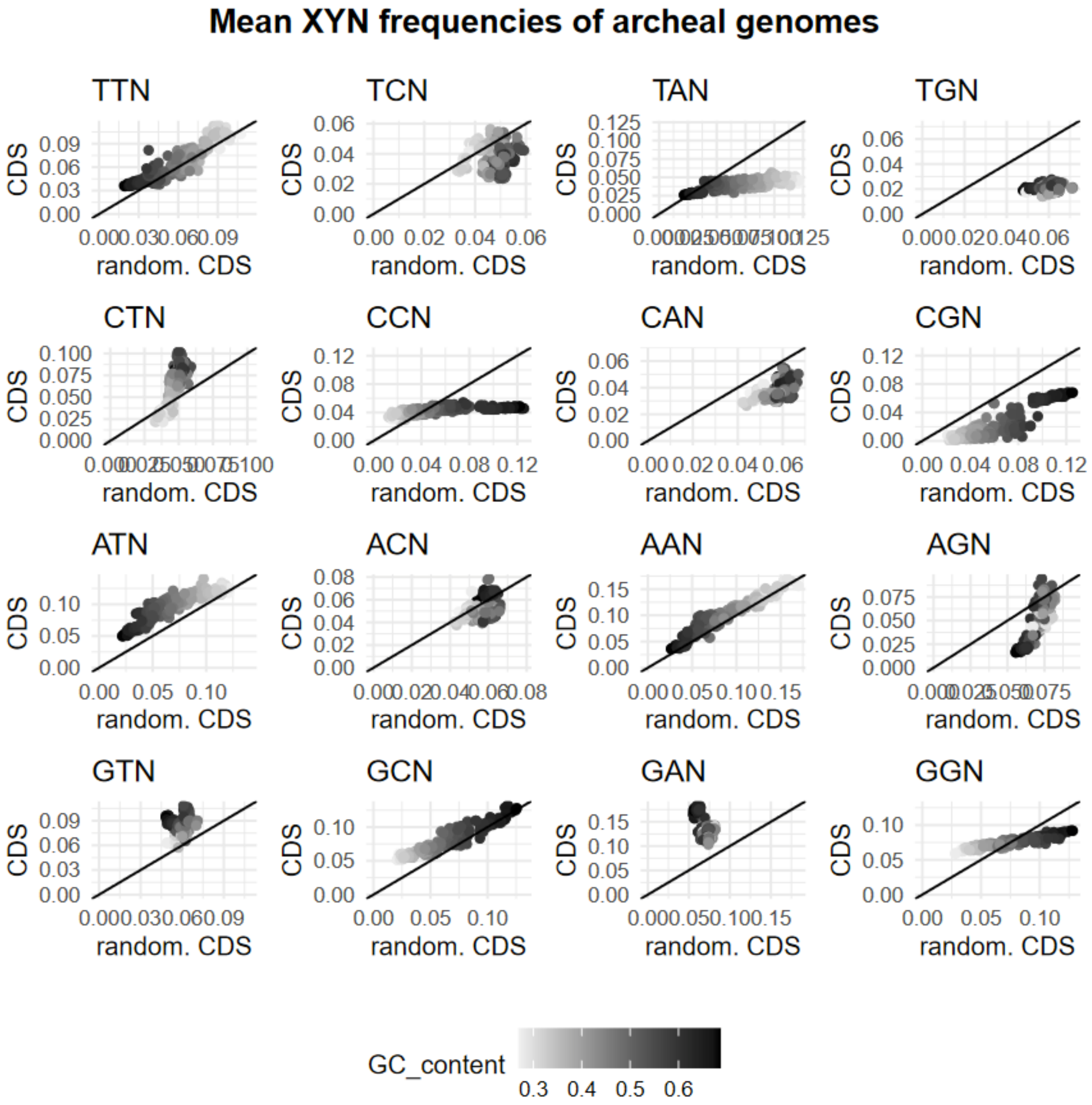


figure 5 : For codons sharing the same first two bases, correlation between the sum of their mean relative frequency among CDS vs this sum in randomized CDS.

$$P\left(X=0\right)=\binom{0}{100}\times\left(3/64\right)^{0}\times\left(1-\left(3/64\right)\right)^{100-0}$$

$= 0.82\%$

- Then, three other codons : the Cys (known as an alpha-helix breaker) encoding codons, and Trp (which is a very large residue, little used in protein structures) codon clearly participate in the 1T pattern.

These last three codons with the opale stop codon also explain the TGN pattern in figure 5. These TGN codons associated with the four CGN Arg-encoding codons largely influence the 2G contrast between CDS and randomized CDS.

This pattern can be translated in the G2 pattern (figure 4) which possesses the smallest CDS range. Interestingly, the codon table is built in such a way that the CDS G1 which presents a complementary tendency with CDS G2, contains many codons that are increased in CDS and four clearly CDS biased amino acids (Val ,Ala, Asp, Glu). As a consequence of these G2 vs G1 opposed tendencies, Gly is very constrained in CDS since its four codons are GGN.

G1 (opposed to G2 and G3) and 1G (opposed to 1T, 1C, and 1A), are notably influenced by the four GAN codons, which are especially increased in CDS. These codons are encoding Asp and Glu which are the two negatively charged amino acids. An hypothesis about this enrichment is that negatively charged amino acids could have a translational efficiency purpose : the exit tunnel of the ribosome being also negatively charged, they could facilitate the translation and exit of the nascent peptide. Placed at the protein surface, they could be a way to avoid non-specific interactions with the ribosome which is the most abundant macromolecule of the cell and with nucleic acids in general [9].

2T pattern reflects a protein structure constraint

Hydrophobic amino acids play a crucial role in the formation of proteins secondary structures [10]. Given the fact that the NTN (2T) column of the codon table contains five hydrophobic amino acids, the bias in this column might impact the protein capacity to fold to a 3D structure (i.e. its foldability potential). In order to verify this hypothesis, we computed the Hydrophobic Cluster Analysis (HCA) score of translated sequences [11]–[14] :

Hydrophobic amino acids clusters of a given sequence are described as associated with regular secondary structures which are characteristic of folded protein domains [15]. Parts of the sequence that link these clusters (referred as linkers) are associated with loops of disordered regions. The sequence of hydrophobic clusters and linkers can be summarized in a foldability score (HCA score).

The NTN (2T) column contains 16 out of the 19 codons encoding the amino acids used to define HCA clusters (the remaining 3 codons encode Tyr and Trp and mostly have low frequencies among CDS). Figure 6 shows that despite the GC content variation, the CDS HCA score is restricted in a small range of fold potential, while random sequences HCA score is very sensitive to the variation of the genome GC content. In addition, for each bin, the CDS IQR is thinner than the random one, emphasizing the idea of a non-GC constraint. This pattern is consistent with the NTN frequency distribution against GC content as represented in figure 7. As the HCA score also depends on the size and number of hydrophobic clusters of a sequence, we verified that these two parameters are also maintained in CDS despite GC content variation (see Supplementary Material). At the end, CDS sequences display a constrained (summed) NTN codons frequency in order to maintain their foldability.

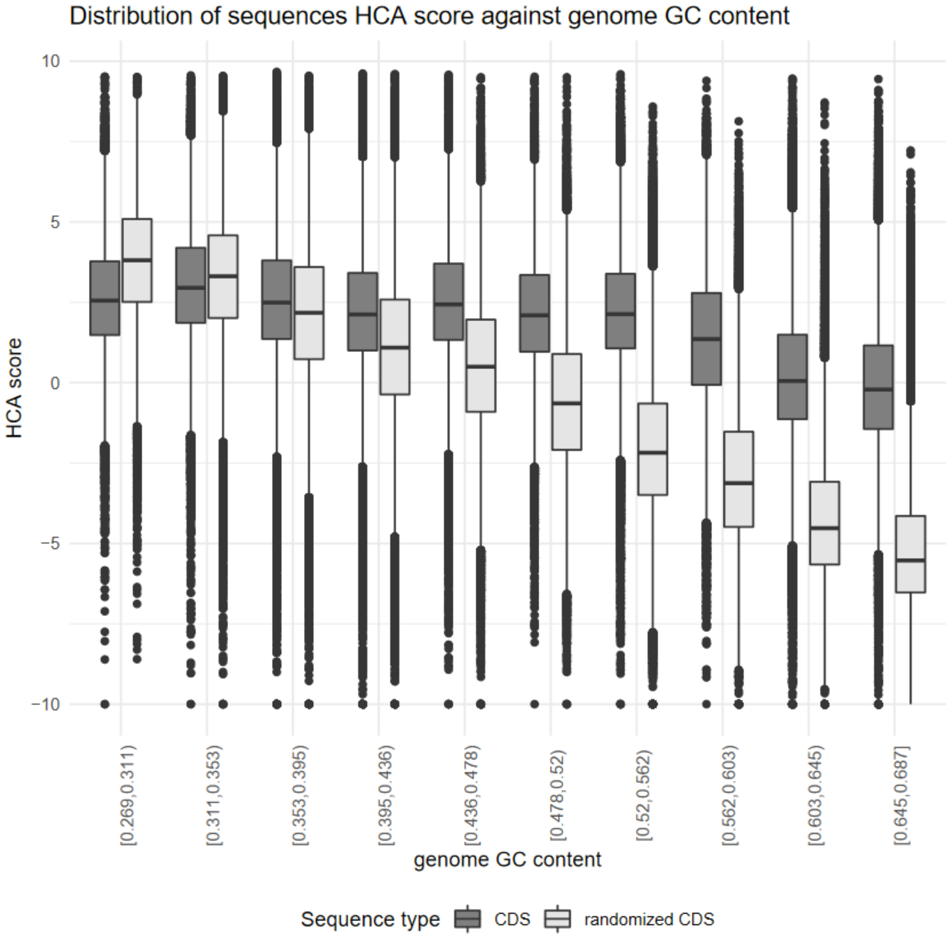


figure 6 : Distribution of the HCA score of sequences for 10 GC content bins.

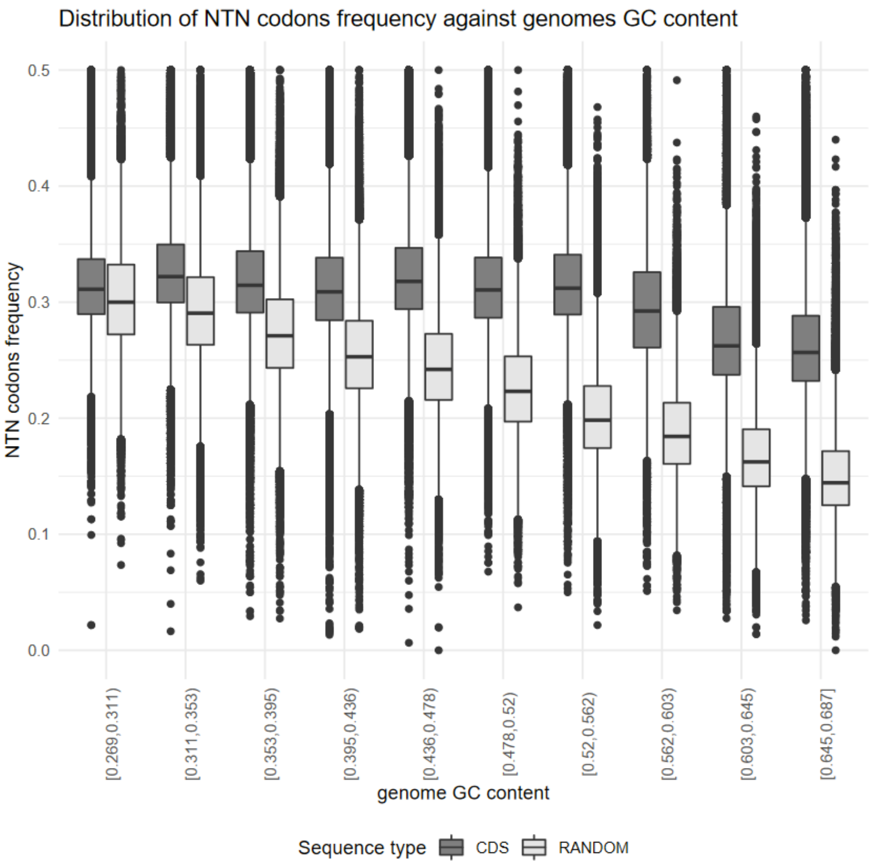


figure 7 : Distribution of the NTN frequency of sequences for 10 GC content bins.

The sum of patterns (CDS vs random CDS) as a coding profile

Taken all together, these patterns seem to define a “coding profile” by contrast with random (figure 8, figure 9). This profile is quite clear at genomes level. But is that profile strong enough to discriminate between a single CDS and a random sequence of same length and base composition ?

The use of an Extreme Gradient Boosting classification model, trained with the previously described inputs : base frequencies, individual codon frequency, XYN count and same-base and same-position count, allows to accurately classify 98.90% of CDS or randomized CDS among Archaea, with one model trained per genome (280 models in total, mean number of sequences : 2461). A similar result is obtained when building 10 models of 28 archeal genomes each of similar GC content (mean bin accuracy : 99.13%), and with a single model built with all the 280 archeal genomes (accuracy : 99.00% (sd 0.42%)). This indicates that despite displaying different CDS vs randomized CDS patterns, these genomes share a similar coding profile once their contrasted codon bias originated in their different GC content is taken into account.

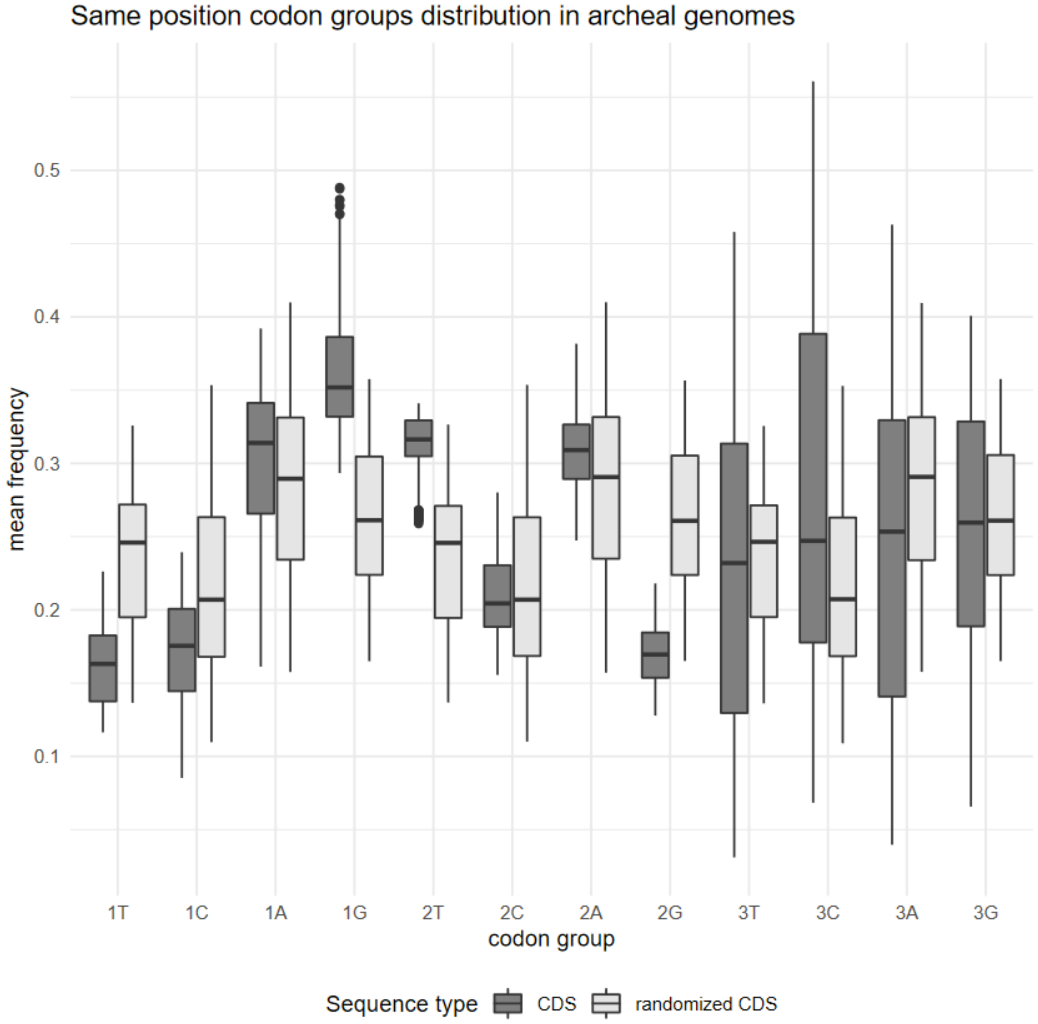


figure 8 : The same-position distribution is an aspect of the coding profile in Archaea CDS.

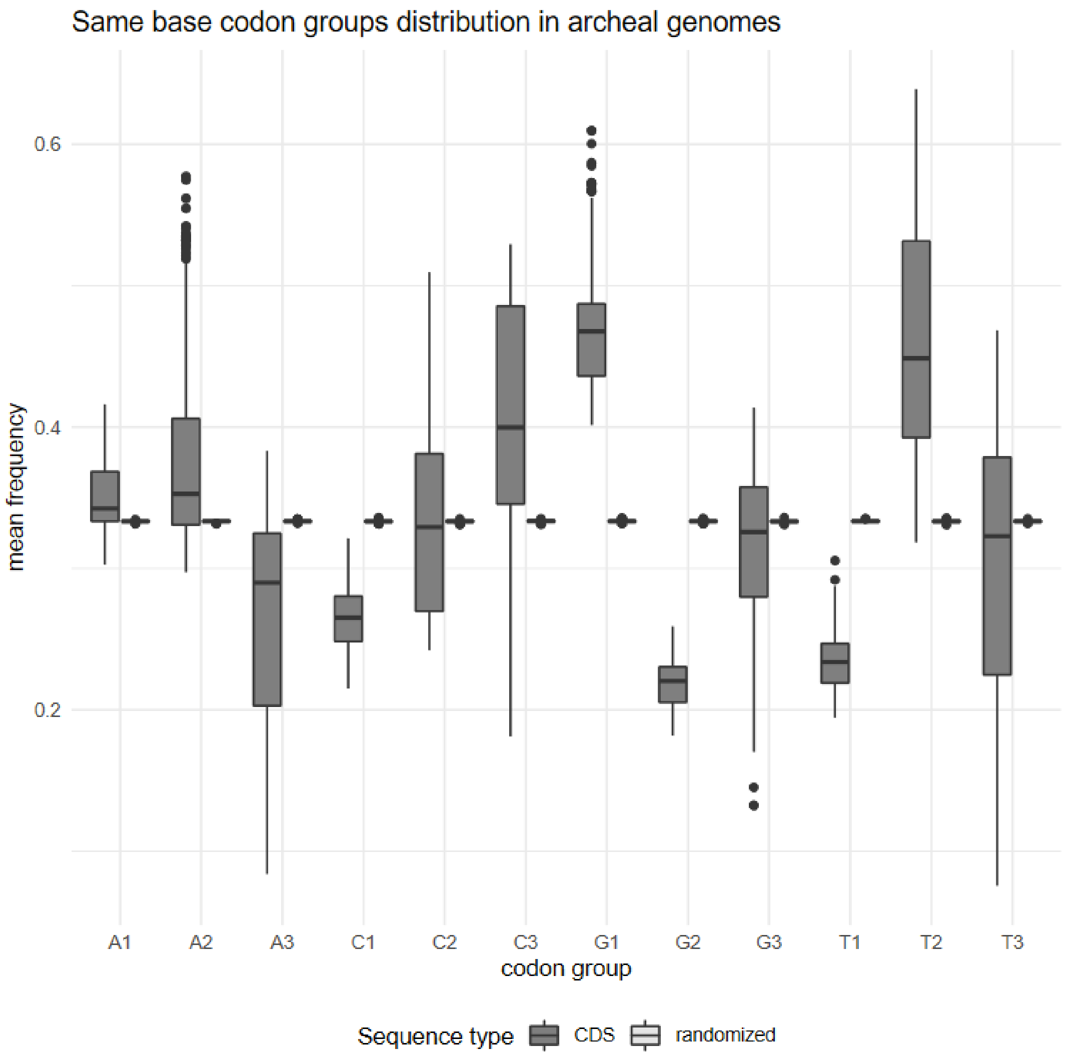


figure 9 : The same-base distribution is an aspect of the coding profile in Archaea CDS.

The coding profile of *S. cerevisiae* is not found in youngest CDS

Additionally, the coding profile described for archaea genomes is extendable to the eukaryote model *Saccharomyces cerevisiae*, which shows pattern similarities with an Archaea genome of equivalent GC content. *S. cerevisiae* genome GC content being about 38%, we tried to predict its CDS or randomized CDS using the model that was built on the 28 archaea of the 36.6-40.1% GC content bin, which produced 87.39% of accurate predictions.

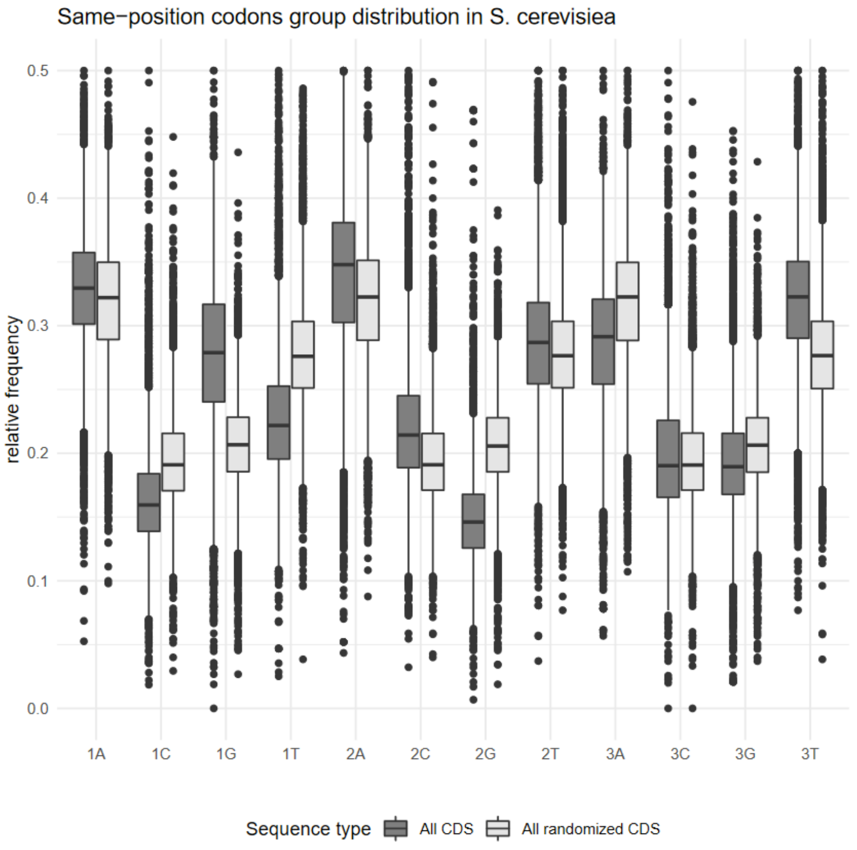


figure 10 : The same-base distribution of *S. cerevisiae* sequences shows that the coding profile is perceptible event at the individual sequences scale.

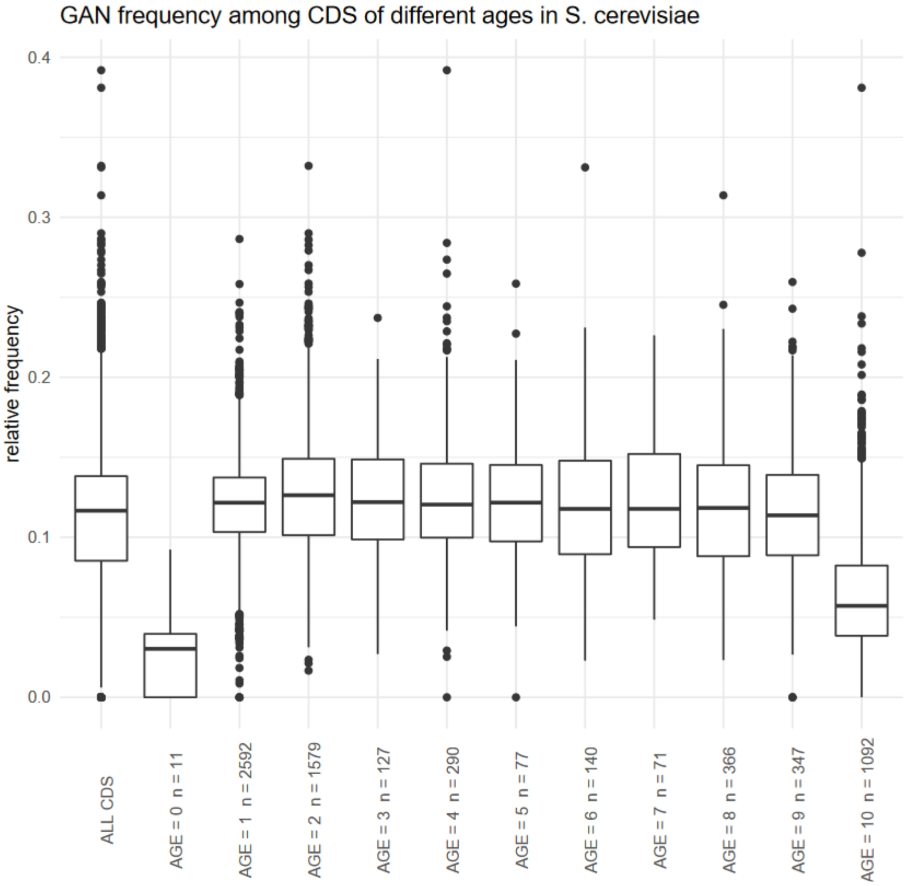


figure 11 : If the other ages distribution is similar to the overall XYN distribution, the XYN distribution of *S. cerevisiae* sequences of age 10 is radically less contrasted. Here is displayed the GAN frequency as an illustration.

The quality of the annotation and the phylogeny of Yeast is the opportunity to look at the influence of gene age on its CDS coding profile. Indeed, if we consider that the age of most *S. cerevisiae* genes can be approximated by the position of the most phylogenetically distant species sharing an orthologue, then we can examine the hypothetical impact of a gene's age on its related coding profile (see Materials and Methods).

Here we segregated the genes of *S. cerevisiae* ranging from 0 (found in more than 10 neighbour species) to 10 (only found in *S. cerevisiae*).

When building a specific classification model for *S. cerevisiae* we obtained 96.11% (sd 0.46%) of accurate predictions. Looking at the misclassified sequences, it appears that 435 out of 560 errors were made on CDS of age 10.
An obvious reason for that is that the youngest CDS are ~~way~~ shorter than older ones, hence lowering the associated statistical power. However, size does not seem to be the only factor of this weaker coding profile among young CDS. More than two third of misclassified CDS of age 10 had a size between 200 and 500 nucleotides (71.4%). For this length range, the length distribution of misclassified sequences is roughly similar between age 10 and the other ages (wilcoxon test p-value : p-value = 0.1075). Among all CDS contained in this length range, 431 out of 2696 sequences were misclassified, while for CDS of age 10 in this range 324 sequences were misclassified out of 1453 sequences, giving a Fisher exact test of p-value of 3.288e-05 (one-sided). In other words, if the youngest CDS do not have yet a clear coding profile because of their size, we can hypothesize that at least some of them have also not been optimized enough by selection at the protein level to acquire such a profile.

Conclusion

GC content alone is far from being sufficient to explain codon frequencies even at the genome level in Archaea and *S. cerevisiae*. By contrast with random sequences of same GC content and length, CDS of these species displays several patterns that have a clear proteic origin. Other non-GC influences might be found through these patterns, from which we could learn more about proteic structural constraints or the CDS shaping in general.
Taken all together, these patterns form a coding profile that makes a CDS distinguishable from a random sequence in a very large majority of cases. However, the youngest CDS of *S. cerevisiae* as a whole do not display a strong coding profile. The study of underlying causes, including their average size, as well as the possibility that they did not endure enough optimization that could be seen through the mentioned patterns is now crucial in order to better understand how new coding sequences arise.

References

[1] U. Lagerkvist, "'Two out of three': an alternative method for codon reading," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 75, no. 4, pp. 1759–1762, Apr. 1978.

[2] A. Şen, K. Kargar, E. Akgün, and M. Ç. Pınar, "Codon optimization: a mathematical programing approach," *Bioinformatics*, vol. 36, no. 13, pp. 4012–4020, Jul. 2020.

[3] S. L. Chen, W. Lee, A. K. Hottes, L. Shapiro, and H. H. McAdams, "Codon usage between genomes is constrained by genome-wide mutational processes," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 10, pp. 3480–3485, Mar. 2004.

[4] S. Zamenhof, G. Brawerman, and E. Chargaff, "On the desoxypentose nucleic acids from several microorganisms," *Biochim. Biophys. Acta*, vol. 9, no. 4, pp. 402–405, Oct. 1952.

[5] i2bc, "GitHub - i2bc/ORFmine." https://github.com/i2bc/ORFmine (accessed Aug. 30, 2021).

[6] "Website." Elek A, Kuzman M, Vlahovicek K (2021). coRdon: Codon Usage Analysis and Prediction of Gene Expressivity. R package version 1.10.0, https://github.com/BioinfoHR/coRdon.

[7] B. A. Wilson, S. G. Foy, R. Neme, and J. Masel, "Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of Gene Birth," *Nat Ecol Evol*, vol. 1, no. 6, pp. 0146–0146, Jun. 2017.

[8] "stats package - RDocumentation." https://rdocumentation.org/packages/stats/versions/3.6.2 (accessed Aug. 30, 2021).

[9] P. E. Schavemaker, W. M. Śmigiel, and B. Poolman, "Ribosome surface properties may impose limits on the nature of the cytoplasmic proteome," *Elife*, vol. 6, Nov. 2017, doi: 10.7554/eLife.30084.

[10] H. J. Dyson, P. E. Wright, and H. A. Scheraga, "The role of hydrophobic interactions in initiation and propagation of protein folding," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 35, pp. 13057–13061, Aug. 2006.

[11] T. Bitard-Feildel and I. Callebaut, "Exploring the dark foldable proteome by considering hydrophobic amino acids topology," *Scientific Reports*, vol. 7, no. 1. 2017. doi: 10.1038/srep41425.

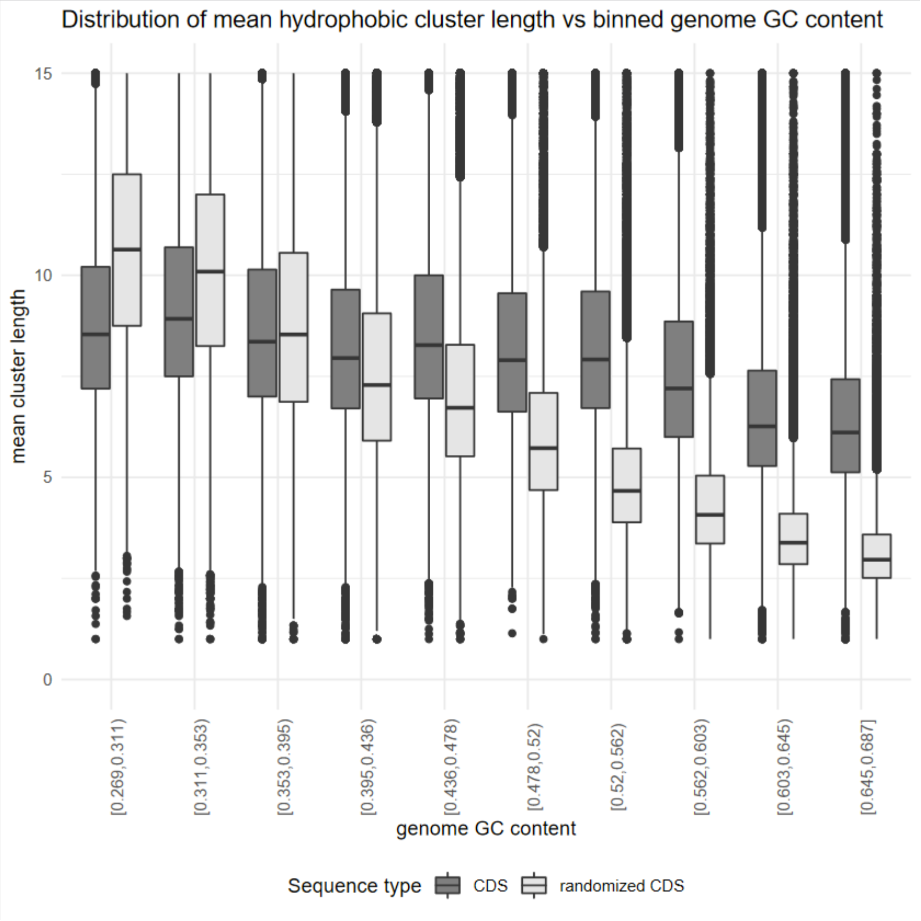
[12] T. Bitard-Feildel, A. Lamiable, J.-P. Mornon, and I. Callebaut, "Order in Disorder as Observed by the 'Hydrophobic Cluster Analysis' of Protein Sequences," *PROTEOMICS*, vol. 18, no. 21–22. p. 1800054, 2018. doi: 10.1002/pmic.201800054.

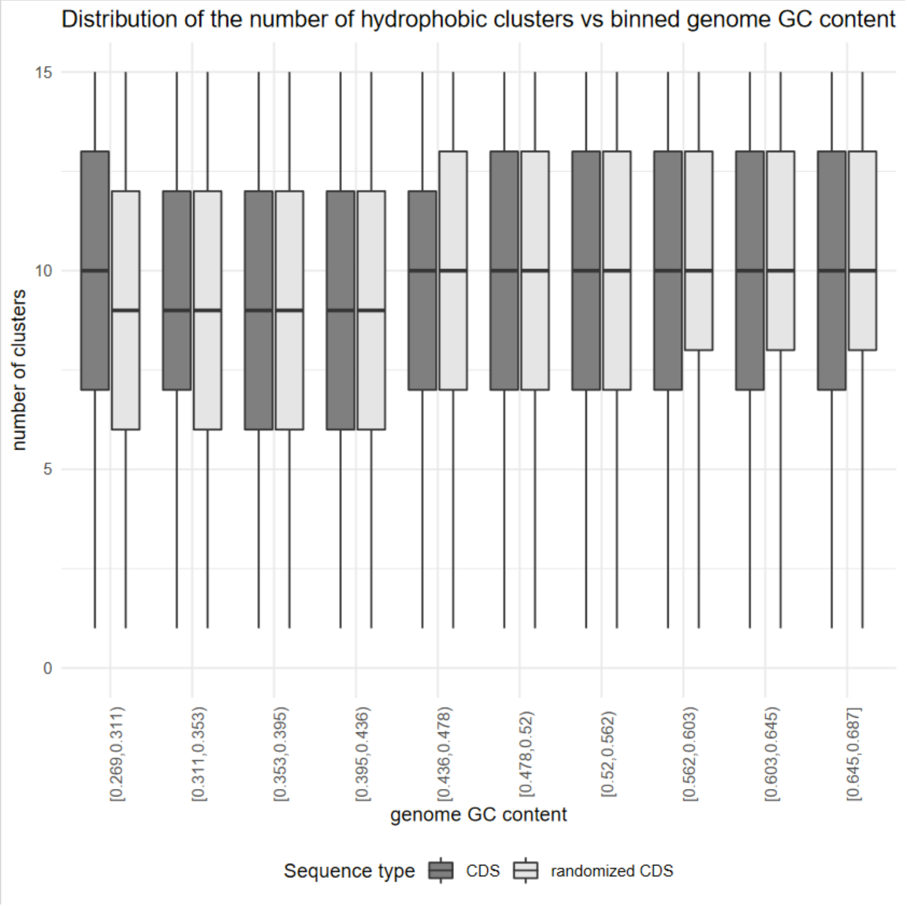
[13] T. Bitard-Feildel and I. Callebaut, "HCAtk and pyHCA: A Toolkit and Python API for the Hydrophobic Cluster Analysis of Protein Sequences." doi: 10.1101/249995.

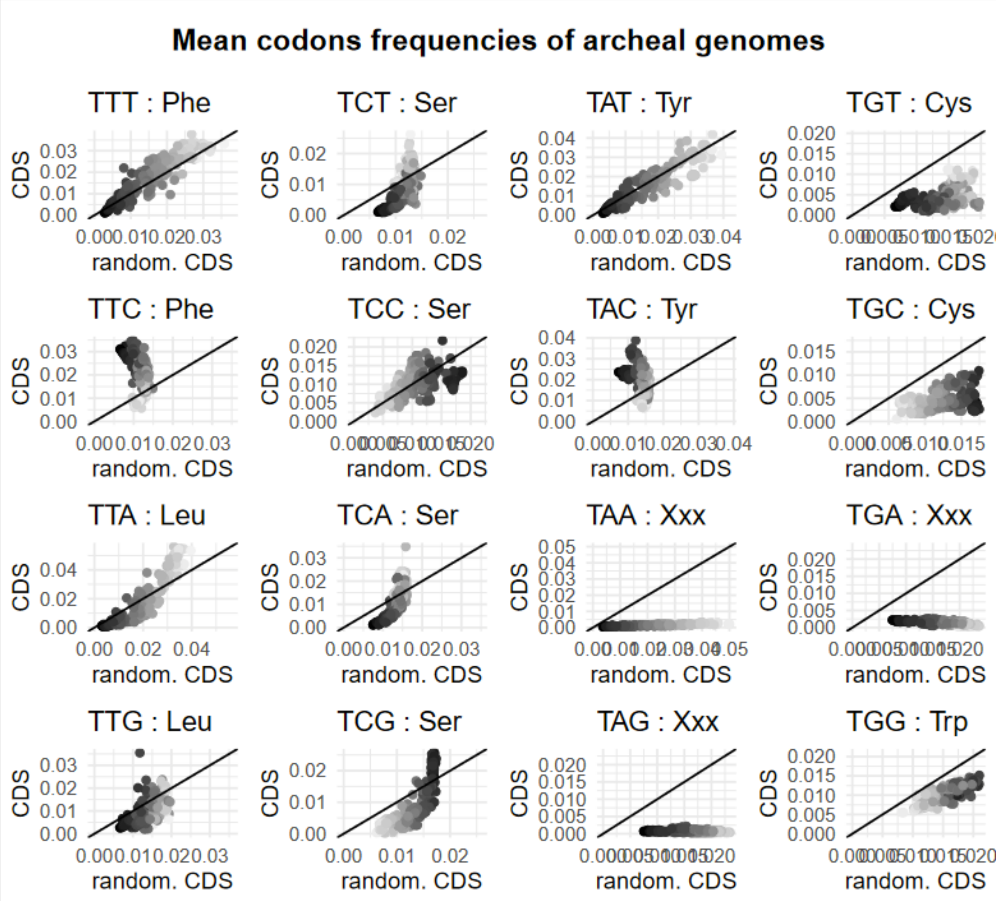
[14] G. Faure and I. Callebaut, "Comprehensive repertoire of foldable regions within whole genomes," *PLoS Comput. Biol.*, vol. 9, no. 10, p. e1003280, Oct. 2013.

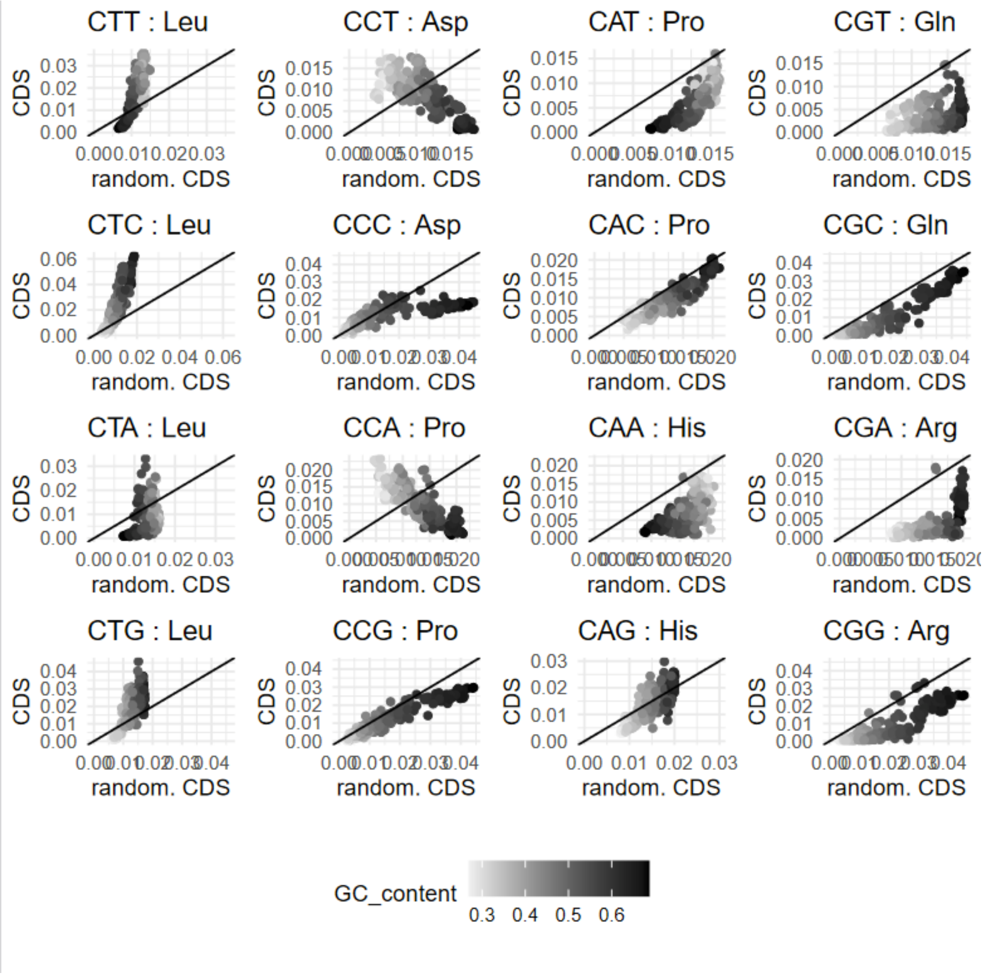
[15] A. Lamiable *et al.*, "A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis," *Biochimie*, vol. 167. pp. 68–80, 2019. doi: 10.1016/j.biochi.2019.09.009.

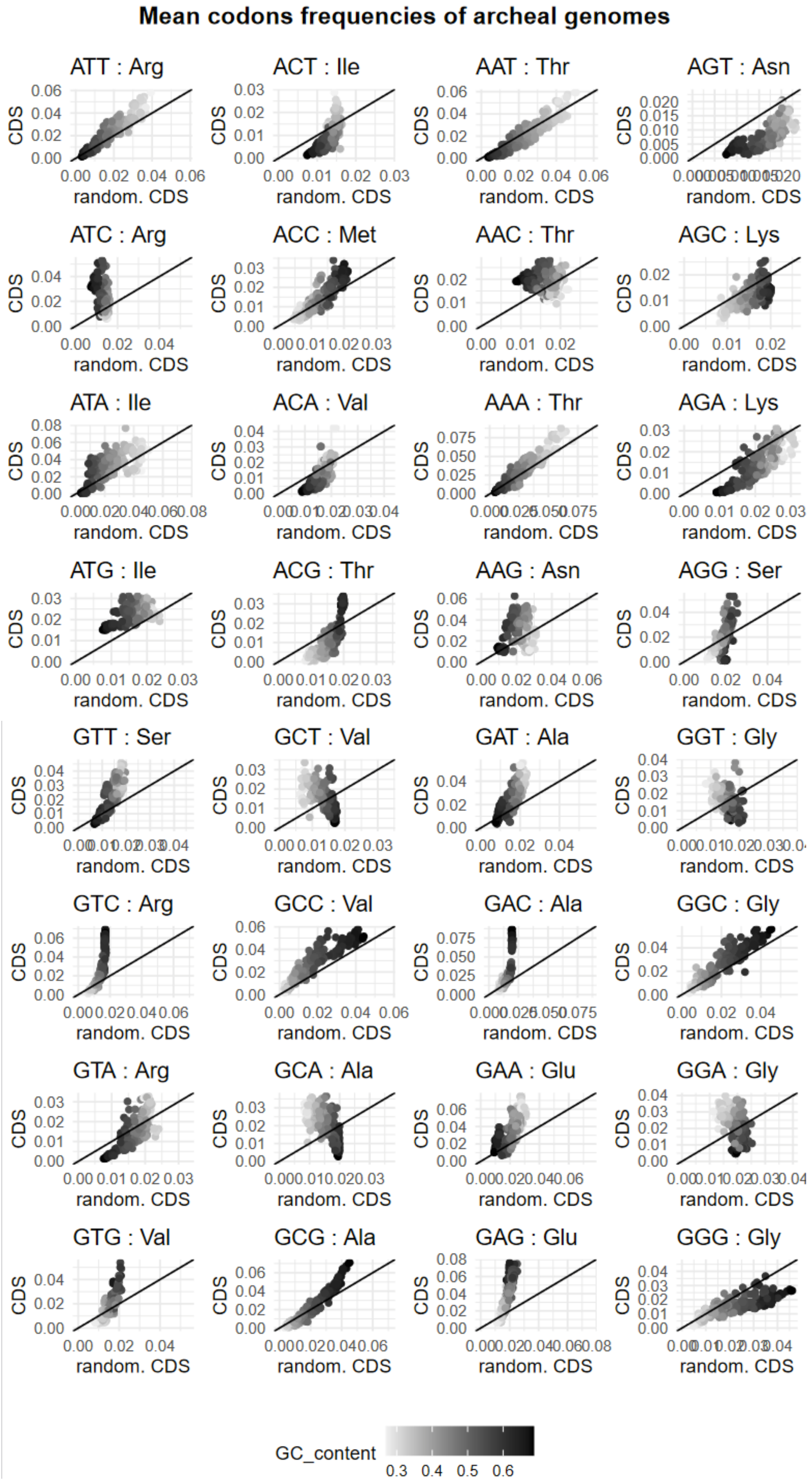
Supplementary Material






Mean codons frequencies of archeal genomes
TTT : Phe
TCT : Ser
TAT : Tyr
TGT : Cys
TTC : Phe
TCC : Ser
TAC : Tyr
TGC : Cys
TTA : Leu
TCA : Ser
TAA : Xxx
TGA : Xxx
TTG : Leu
TCG : Ser
TAG : Xxx
TGG : Trp
CDS
random. CDS


CTT : Leu
CCT : Asp
CAT : Pro
CGT : Gln
CTC : Leu
CCC : Asp
CAC : Pro
CGC : Gln
CTA : Leu
CCA : Pro
CAA : His
CGA : Arg
CTG : Leu
CCG : Pro
CAG : His
CGG : Arg
CDS
random. CDS
GC_content
0.3
0.4
0.5
0.6


Mean codons frequencies of archeal genomes
ATT : Arg
ACT : Ile
AAT : Thr
AGT : Asn
ATC : Arg
ACC : Met
AAC : Thr
AGC : Lys
ATA : Ile
ACA : Val
AAA : Thr
AGA : Lys
ATG : Ile
ACG : Thr
AAG : Asn
AGG : Ser
GTT : Ser
GCT : Val
GAT : Ala
GGT : Gly
GTC : Arg
GCC : Val
GAC : Ala
GGC : Gly
GTA : Arg
GCA : Ala
GAA : Glu
GGA : Gly
GTG : Val
GCG : Ala
GAG : Glu
GGG : Gly
CDS
random. CDS
GC_content
0.3
0.4
0.5
0.6