ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

www.sciencedirect.com/science/journal/16720229

GENOMICS
PROTEOMICS &
BIOINFORMATICS

Article

On the Organizational Dynamics of the Genetic Code

Zhang Zhang[1] and Jun Yu[1,2*]

[1]*Plant Stress Genomics Research Center, Division of Chemical and Life Sciences and Engineering, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia;*
[2]*CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China.*

Genomics Proteomics Bioinformatics 2011 Apr; 9(1-2): 21-29 DOI: 10.1016/S1672-0229(11)60004-1

Received: Aug 30, 2010; Accepted: Oct 26, 2010

Abstract

The organization of the canonical genetic code needs to be thoroughly illuminated. Here we reorder the four nucleotides—adenine, thymine, guanine and cytosine—according to their emergence in evolution, and apply the organizational rules to devising an algebraic representation for the canonical genetic code. Under a framework of the devised code, we quantify codon and amino acid usages from a large collection of 917 prokaryotic genome sequences, and associate the usages with its intrinsic structure and classification schemes as well as amino acid physicochemical properties. Our results show that the algebraic representation of the code is structurally equivalent to a content-centric organization of the code and that codon and amino acid usages under different classification schemes were correlated closely with GC content, implying a set of rules governing composition dynamics across a wide variety of prokaryotic genome sequences. These results also indicate that codons and amino acids are not randomly allocated in the code, where the six-fold degenerate codons and their amino acids have important balancing roles for error minimization. Therefore, the content-centric code is of great usefulness in deciphering its hitherto unknown regularities as well as the dynamics of nucleotide, codon, and amino acid compositions.

Key words: genetic code, codon, GC content, purine content, organizational dynamics, compositional dynamics

Introduction

The canonical genetic code encodes 20 amino acids (as well as the start and stop signals) redundantly by its 64 triplet codons as combinations of the four nucleotides, thymine (T), cytosine (C), adenine (A) and guanine (G). Obviously, codons and amino acids are not randomly associated, and it is proposed to be systematically related to the origin and evolution of the genetic code (*1-7*) and the physicochemical properties of the 20 amino acids (*8-11*). Therefore, deciphering the relationship of the codons and amino acids in the genetic code is of great significance, not only in better understanding the code but also in providing insights into evolutionary mechanisms of DNA sequences among organisms (*12-15*).

In a large variety of publically available genomes, codons and amino acids are not used randomly. A number of studies have been conducted to investigate this non-randomness based on the genetic code, which is organized traditionally by ordering four nucleotides as T, C, A, G. In contrast, it is argued by recent studies that the genetic code is more appropriate to be reorganized based on alternative nucleotide orders (*16*) or contents (*17, 18*). A useful proposal is the content-centric reorganization of the genetic code based

*Corresponding author.
E-mail: junyu@big.ac.cn

on GC (guanine plus cytosine; G+C) and purines (adenine plus guanine; A+G or R). The content-centric genetic code promises to explain intrinsic relationship between protein-coding sequences and codon/amino acid compositions (*17*). However, little attention has been paid enough to study the compositional dynamics within such a content-centric genetic code. Therefore, the purpose of this study is to decipher the underlying patterns of the genetic code through a quantitative analysis of codon and amino acid usages. We provide an algebraic representation for the content-centric genetic code, and mathematically demonstrate its classification schemes based on GC and purine contents. Based on a large collection of 917 prokaryotic genomes, we quantify codon and amino acid usages, relate the usages to the intrinsic organization of the content-centric genetic code, and explore the usages under different classification schemes. We further investigate the non-random allocation of codons/amino acids in the genetic code, uncover the potential roles of six-fold degenerate codons and finally provide in-depth discussions on the balance of nucleotide content and physicochemical properties in the genetic code.

Results and Discussion

An algebraic representation of the genetic code

The canonical genetic code is composed of 64 triplets from the permutation of T, C, A, G. It is speculated that the triplet code evolves from a doublet code (*18-23*) and that A and T are believed to be more ancient than G and C according to their chemical properties (*24-26*). Therefore, according to their emergence (*27, 28*), we reorder the four nucleotides as A, T, G, C (unlike the traditional order T, C, A, G). A vector of the four nucleotides is then defined as V=[A T G C], and thus its transpose is:

$$V^{T}=\begin{bmatrix}A\\T\\G\\C\end{bmatrix}$$

The genetic code can be represented algebraically as a three-dimensional matrix where each dimension represents one of the three positions in the triplet code. It is well established that the first and second codon positions have a crucial role in determining the structure of the genetic code (*29, 30*). Therefore, we first construct a doublet code as:

$$D=V^{T}\times V=\begin{bmatrix}A\\T\\G\\C\end{bmatrix}\times\begin{bmatrix}A & T & G & C\end{bmatrix}=\begin{bmatrix}AA & AT & AG & AC\\TA & TT & TG & TC\\GA & GT & GG & GC\\CA & CT & CG & CC\end{bmatrix} \qquad (1)$$

Hence, a triplet code representing the genetic code can be built based on D:

$$X=DN=\begin{bmatrix}AAN & ATN & AGN & ACN\\TAN & TTN & TGN & TCN\\GAN & GTN & GGN & GCN\\CAN & CTN & CGN & CCN\end{bmatrix} \qquad (2)$$

or

$$X^{T}=D^{T}N=\begin{bmatrix}AAN & TAN & GAN & CAN\\ATN & TTN & GTN & CTN\\AGN & TGN & GGN & CGN\\ACN & TCN & GCN & CCN\end{bmatrix} \qquad (3)$$

where N is one of the four nucleotides.

Suppose that S=G or C, and W=A or T, then X^T can also be reformatted concisely as:

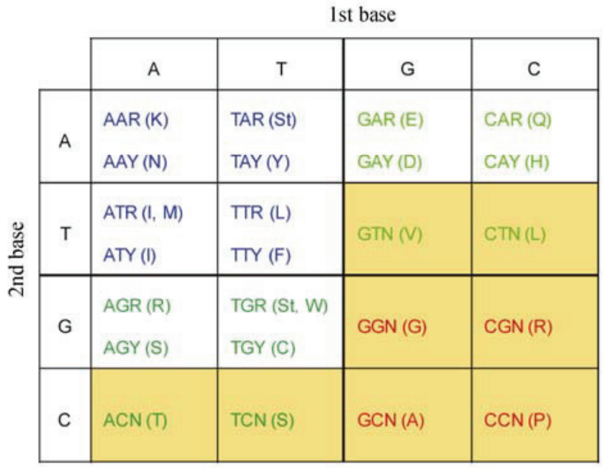
$$X^{T}=\begin{bmatrix}WW & SW\\WS & SS\end{bmatrix}N \qquad (4)$$

where

$$WW=\begin{bmatrix}AA & TA\\AT & TT\end{bmatrix} \quad SW=\begin{bmatrix}GA & CA\\GT & CT\end{bmatrix}$$

$$WS=\begin{bmatrix}AG & TG\\AC & TC\end{bmatrix} \quad SS=\begin{bmatrix}GG & CG\\GC & CC\end{bmatrix}$$

Unlike earlier attempts of algebraic representations of the genetic code (*5, 31-34*), X (or X^T) is organized as a doublet code D (or D^T) appending nucleotide N, which is based on the assumption that the triplet code evolves from the doublet code (*6, 18-22*) and the first two positions have a determinative role in the structure of the genetic code (*29, 30*). The algebraic representation X (or X^T) is essentially equivalent to a content-centric organization of the genetic code, as proposed previously (*17*). Based on X^T, the genetic code is depicted as shown in **Figure 1**.



2nd base \ 1st base	A	T	G	C
A	AAR (K) AAY (N)	TAR (St) TAY (Y)	GAR (E) GAY (D)	CAR (Q) CAY (H)
T	ATR (I, M) ATY (I)	TTR (L) TTY (F)	GTN (V)	CTN (L)
G	AGR (R) AGY (S)	TGR (St, W) TGY (C)	GGN (G)	CGN (R)
C	ACN (T)	TCN (S)	GCN (A)	CCN (P)

Figure 1 Illustration of the genetic code based on an algebraic representation X^T. Codons with yellow background encode the same amino acid, independent of the third base.

Two halves of the genetic code

As shown in X^T and Figure 1, we found that the genetic code is clearly divided into two halves with distinct features. One half includes eight robust doublets (AC, TC, GT, CT, GG, CG, GC and CC in yellow background) at codon positions 1 and 2 (cp1 and cp2) and N, standing for any four nucleotides, at codon position 3 (cp3; *e.g.*, all four codons associated with ACN encode Thr); therefore, they are not sensitive to CG content changes at cp3. We termed this half as the pro-robustness half (PRH), including 32 codons and 8 amino acids. Conversely, the other half is very sensitive to purine changes at cp3; only when there is a purine (A or G, denoted as R) or a pyrimidine (T or C, denoted as Y) each encodes the same amino acid (*e.g.*, AAR codes for Lys, and AAY for Asn), with the exception of the two doublets, AT (ATA for Ile and ATG for Met) and TG (TGA for stop and TGG for Trp). This half contains 32 codons and 15 amino acids (three amino acids, Ser, Arg and Leu, with the highest level of codon degeneracy, are distributed in both of the two halves) as well as three stop and one start signals, so that we denoted this half as the pro-diversity half (PDH).

According to the two halves (PDH and PRH), we observed that GC content offers robustness, whereas purine content supports diversity (*17*). As GC content varies from 20% to 80%, codon usages change significantly. Accordingly, incorporation of GC content into the reorganization of the genetic code provides a clearer illustration on diversity and robustness. In addition to GC content, however, there may be other crucial factor(s); evidence has accumulated that purines have an important role in determining amino acid physicochemical properties (*9, 10, 30*) (described below). Moreover, it is notable that all doublets in PRH are purine-insensitive, whereas most doublets in PDH are purine-sensitive. As compared to GC content, purine content fluctuates narrowly from ~40% to ~60% in a total of 917 prokaryotic genome sequences, which also reflects diverse interplays of mutation and selection acting on different genomes (*35, 36*). Therefore, X^T, a reorganization based on GC and purine contents, promises to capture more features underlying the genetic code.

The classification of PDH and PRH indicates the possible role of nucleotide content in determining codon usage. To investigate this possibility, we examined the relationship between GC content and total frequencies of 32 codons locating in PDH and PRH, respectively (**Figure 2**). Based on a collection of 917 prokaryotic genome sequences, we found that GC content exhibits a significant correlation with the total codon usage in each half: negative in PDH and

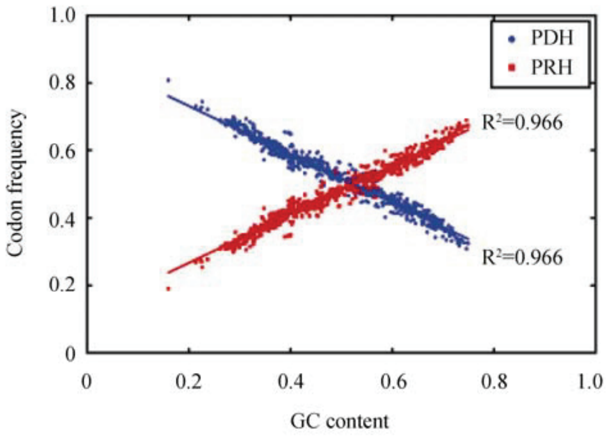


Figure 2 Correlation between GC content and codon frequencies in PDH and PRH, based on a variety of prokaryotic genome sequences. Each point represents a genome sequence. The linear regression results with squared correlation coefficient (R^2) are $y=-0.719x+0.877$ (R^2=0.966) in PDH and $y=0.719x+0.123$ (R^2=0.966) in PRH, respectively, with two-tailed significance level of $P<0.0001$.

positive in PRH, with both squared correlation coefficients R^2=0.966. Consistent with the expectation, we observed that codon usage in PDH decreases when GC content increases as compared to what in PRH runs toward the opposite direction. As shown in Figure 2, the linear regression lines for PDH and PRH intersect at GC content ≈0.5, indicating equal usage of codons between the two halves. The significant correlations for PDH and PRH suggest that GC content indeed has a determinative role in codon usage in the two halves, which further strengthens the idea that codon usage can be largely inferred from GC content (*37*).

Four quarters of the genetic code

According to the variability and position of GC content, the genetic code can also be divided into four quarters (*17*), in which GC content changes occur (1) at neither cp1 nor cp2 (WWN; AT-rich quarter), (2) only at cp1 (SWN; GCp1 quarter), (3) only at cp2 (WSN; GCp2 quarter), and (4) at both cp1 and cp2 (SSN; GC-rich quarter) (see Equation 4 and Figure 1). With sixteen codons residing in each quarter, the AT-rich quarter encodes seven amino acids (Lys, Asn, Tyr, Ile, Met, Leu and Phe) as well as two stop and one start signals. In Xiao and Yu (*28*), this quarter is proposed to be the core group for diversity in the primordial genetic code. But Hartman (*4*) took a contrasting view that the primordial code is assumed to be a GC code. The GCp1 quarter has six amino acids (Glu, Asp, Gln, His, Val and Leu) and the GCp2 quarter contains five amino acids (Arg, Ser, Trp, Cys and Thr). The GC-rich quarter possesses only four amino acids (Gly, Arg, Ala and Pro) and this quarter is thought to be new comers to the genetic code except Arg that is six-fold degenerate and plays unique roles.

The four-quarter classification scheme provides a clear way in better understanding the compositional dynamics across a variety of species. For example, there is a widely reported phenomenon that GC content at cp1 (denoted as GC1) is often greater than that at cp2 (GC2). To examine this phenomenon under this classification scheme, we first explore whether GC content has any relationship with total usage of 16 codons in each quarter. Across a collection of 917 prokaryotic genome sequences, we found that the total codon usage in the GCp2 quarter never exceeds that in the GCp1 quarter (**Figure 3**), consequently contributing to GC2<GC1 (since the AT-rich and GC-rich quarters have an equal usage of GC content at the first two positions) and consistent well with a previous study (*38*). Interestingly, it is observed for the first time that the total codon usage in the GCp2 quarter also tends to be less than that in the AT-rich quarter, particularly at small values of GC content. In addition, it is also notable that the total codon usage in the GCp2 quarter has no significant correlation with GC content and appears nearly constant across a wide range of GC content (slope=−0.057, R^2=0.309). Conversely, the total codon usages in the rest three quarters correlate significantly with GC content: negative correlation in the AT-rich quarter (R^2=0.958), and positive correlations in the GCp1 (R^2=0.797) and GC-rich quarters (R^2=0.943). These results strongly indicate that for a wide variety of compositions in the prokaryotic genomes, the GCp2 quarter most likely has a special way to maintain its total codon usage at a nearly constant level and always keeps it under-utilized compared to that in the AT-rich and GCp1 quarters. The reasons are three folds. First, Cys and Trp are the least used amino acids among eubacterial genomes. Second, it has six of its sibling codons encoding two of the three amino acids with six-fold

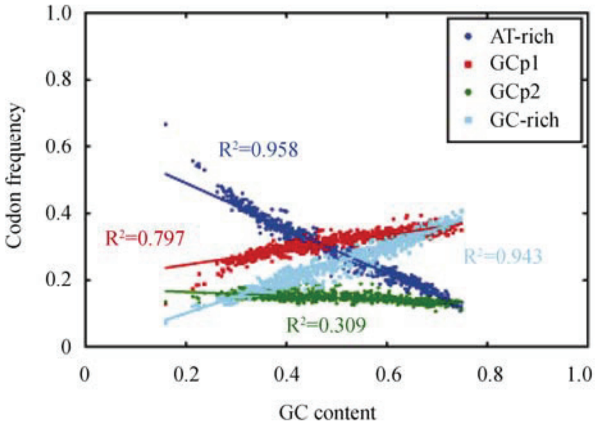


Figure 3 Correlation between GC content and codon frequencies in four quarters (AT-rich, GCp1, GCp2 and GC-rich). Each point represents a genome sequence. The linear regression results with squared correlation coefficients (R^2) are y=−0.674x+0.625 (R^2=0.958) for the AT-rich quarter, y=0.224x+0.201 (R^2=0.797) for the GCp1 quarter, y=−0.057x+0.175 (R^2=0.309) for the GCp2 quarter, and y=0.507x−0.001 (R^2=0.943) for the GC-rich quarter, respectively, with two-tailed significance level of *P*<0.0001.

degenerate codons, and both Ser and Arg are not the most abundant amino acids as compared to Leu. Third, there is a stop codon in this quarter but not in the GCp1 quarter.

Following the nearly constant usage of WSN codons in the GCp2 quarter, we raised a question: “Is it a result of constant usage of each individual WSN codon, or a compensation balance of their inconstant usages?” Based on our collected sequences, WSN codons do not exhibit constant trends with varying GC content, with the only exception of TGG. Considering the third codon position, one half of the GCp2 quarter (AGN and TGN) is sensitive to purines, whereas the other (ACN and TCN) is insensitive to purines. Therefore, attention should be paid to the fact that the sibling codon of TGG is a stop signal (TGA). Hence, we listed AGR (Arg), AGY (Ser), TGY (Cys), ACN (Thr) and TCN (Ser), as sibling codons in this quarter, and estimated the total usage of each sibling codon in our collected genomes. Although AGY and TCN encode for the same amino acid (Ser), we argue that they are not sibling, since their first two nucleotides are different and they may undertake different pathways for composition dynamics (described below). Results show that the sibling codons as well as a single codon TGG tend to be used at nearly constant frequencies across a wide range of GC content, yielding nearly no correlation with GC content, as indicated by their very low squared correlation coefficients (R^2), especially for TGY (R^2=0.007) (**Figure 4**). Linear regression analysis performed in this study also estimates the slope for each regression line, a sign of the sensitivity of one variable to the other. Consistent with low R^2, the absolute slopes for all sibling codons in the GCp2 quarter appear very small, with the upper at 0.048 by AGR and the lower at 0.002 by TGY, revealing extreme insensitivity to GC content variation. According to the above analysis, we conclude that the nearly constant usage of codons in the GCp2 quarter stems from insensitivity of its sibling codons to GC content variation.

Non-random allocation of codons and amino acids

The 64 codons are not randomly allocated in the genetic code (*39*). It can be seen from Figure 1 that: (1) all four-fold degenerate codons locate in PRH and the rest are in PDH; (2) the stop and start signals are all in PDH and the AT-rich quarter contains both stop and start signals; (3) three amino acids with six-fold degenerate codons (Ser, Leu and Arg) are distributed across PDH and PRH and among all four quarters. In detail, their four-fold degenerate codons are high-GC in PRH, whereas their two-fold degenerate codons are low-GC in PDH. Although it seems that they are assigned in a disordered way, these three amino acids are most likely to be selected for balancing GC content between the AT-rich (TTG) and GCp1 (CTN) quarters by Leu, between the GCp2 (AGR) and GC-rich (CGN) quarters by Arg, and within the GCp2 quarter but across PDH (AGY) and PRH (TCN) by Ser. Considering that there are three scenarios coupled with this balance, *i.e.*, unchanged, increased and decreased GC content, this leaves us wondering whether these three amino acids are separately responsible for the three scenarios with no, positive and negative correlations with GC content variation. To test this idea, we plotted frequencies of their two-fold and four-fold degenerate codons separately, as well as their total frequencies for each collected sequence in **Figure 5**. Consistent with our expectations, all three two-fold degenerate codons (TTG, AGR and AGY) correlate negatively with GC content, whereas the two four-fold degenerate codons (CTN and CGN) correlate positively with GC content, with one exception for TCN. Their total usages, *viz.*, amino acid usages, however, present different correlations with GC content. Across a wide variation of GC content, Leu is used with nearly constant frequency (slope=0.021, R^2=0.114) (Figure 5B), whereas Arg and Ser correlate positively (slope=0.116, R^2=0.893) (Figure 5A) and slight negatively (slope=–0.034, R^2=0.394) (Figure 5C) with GC content, respectively, although AGY and TCN (coding for Ser) have very low correlation coefficients and very small slopes. These results indicate that these three amino acids play different roles in balancing GC content: Leu is preferentially used against changing GC content, whereas Arg and Ser are selected for increasing and decreasing GC content, respectively.

The distribution of 20 amino acids in the genetic code is also not random. For any sense codon, the

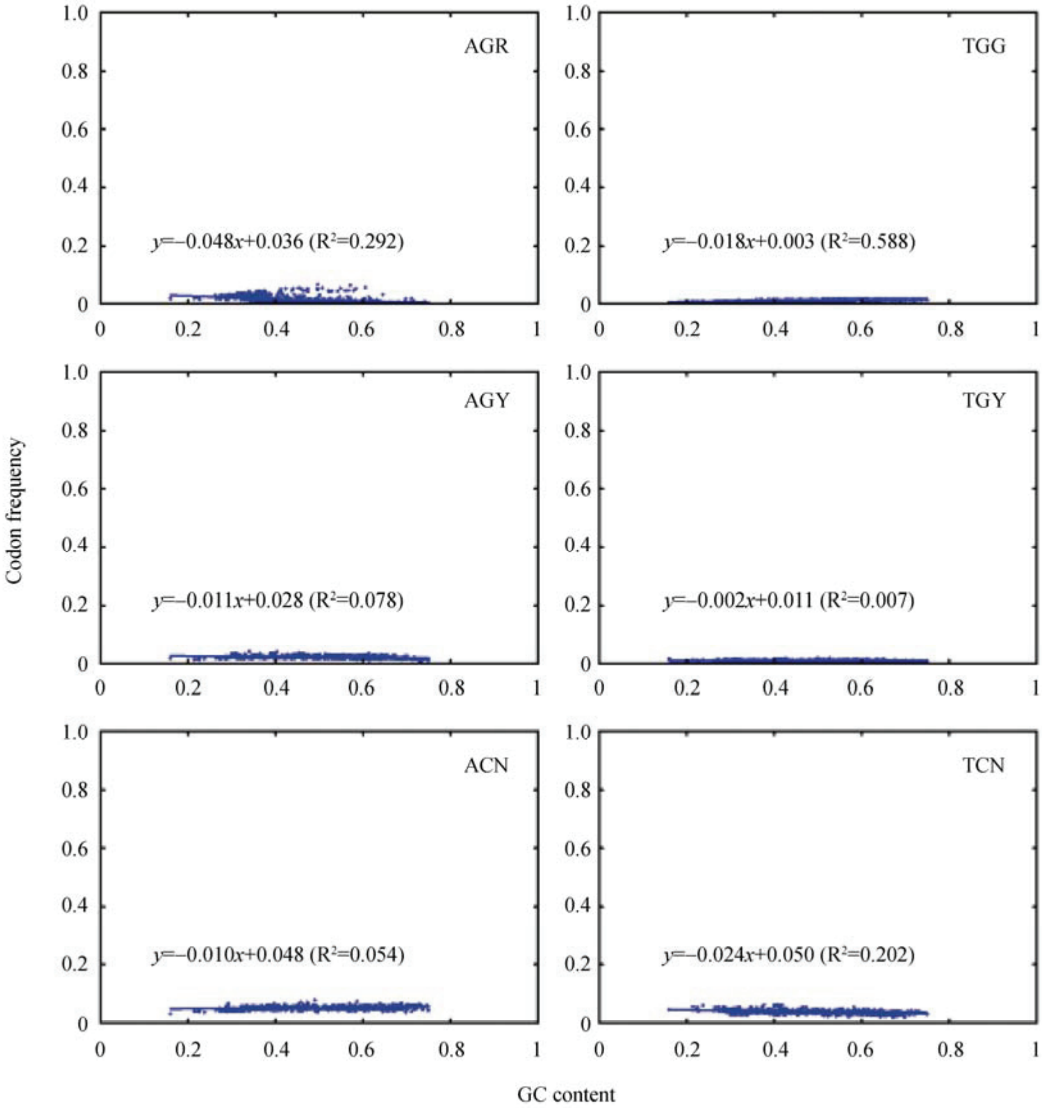


Figure 4 Correlation between GC content and frequencies of sibling codons in the GCp2 quarter. Each point represents a genome sequence. The linear regression lines as well as their corresponding squared correlation coefficients (R^2) are shown in each panel, with two-tailed significance level of P<0.0001.

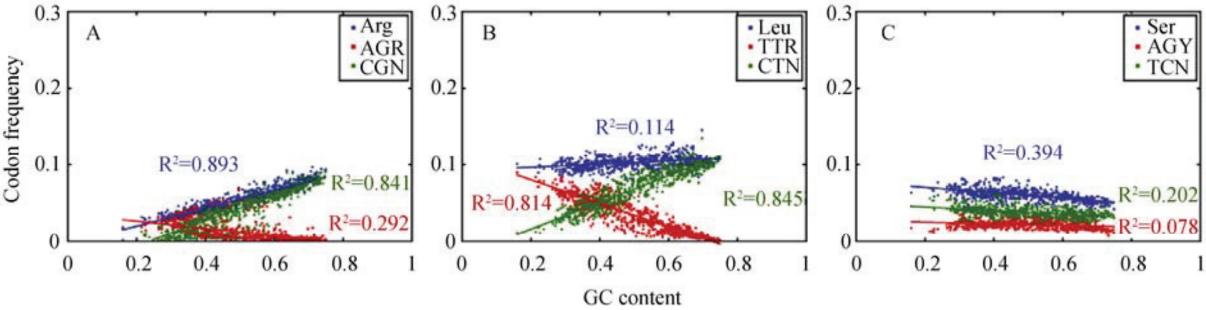


Figure 5 Correlation between GC content and frequencies of three six-fold degenerate amino acids (Arg, Leu and Ser) as well as their two-fold and four-fold degenerate codons. Each point represents a genome sequence. The linear regression results with squared correlation coefficient (R^2) are: Arg: y=0.116x−0.003 (R^2=0.893), AGR: y=−0.048x+0.036 (R^2=0.292), CGN: y=0.165−0.039 (R^2=0.841) (**A**); Leu: y=0.021x+0.092 (R^2=0.114), TTR: y=−0.153x+0.111 (R^2=0.814), CTN: y=0.174x−0.019 (R^2=0.845) (**B**); Ser: y=−0.034x+0.077 (R^2=0.394), AGY: y=−0.011x+0.028 (R^2=0.078), TCN: y=−0.024+0.050 (R^2=0.202) (**C**). The corresponding two-tailed significance levels of P-value are less than 0.0001.

second nucleotide preferentially controls physicochemical properties of its encoding amino acid (*15, 40, 41*). As shown in Figure 1, there is a clear separation of amino acids with similar physicochemical properties into codons with the same nucleotide at cp2 (*30*): (1) Codons NAN contain exclusively polar amino acids (Lys, Asn, Tyr, Glu, Asp, Gln and His; the polar row). (2) Codons NTN possess entirely hydrophobic amino acids (Ile, Met, Leu, Phe and Val; the hydrophobic row). (3) Codons NCN include all small amino acids (Thr, Ser, Ala and Pro; the tiny row). (4) For codons NGN, however, there is no single physicochemical property shared among all encoded amino acids: Trp and Cys are hydrophobic; Arg, Ser, Trp and Cys are polar; Cys, Ser and Gly are small; and Arg and Trp are big. Therefore, we named codons NGN as the mixed row in Figure 1. Additionally, all charged amino acids, including positive (AAR for Lys, CAY for His, and AGR and CGN for Arg) and negative (GAR for Glu, GAY for Asp), are assigned into NRN, suggesting that the charge is preferentially determined by codons with purines at the second position. Similarly, NYN-containing rows are populated with either hydrophobic or polar amino acids but not the charged. Moreover, all charged amino acids are distributed into the four quarters and across PDH and PRH. The non-random distribution of amino acids can be easily explained by selection to minimize deleterious effects of translation errors on physicochemical properties (*42*).

As mentioned above, the three amino acids (Arg, Leu and Ser) with six-fold degenerate codons have different roles in balancing GC content. At the level of amino acid, do they also perform balances for physicochemical properties? If yes, the fundamental requirement of these amino acids is that they should have completely distinct physicochemical properties, so that they can work at different cases for error minimization. Given the three amino acids, there are accordingly three scenarios for each individual physicochemical property (if quantitative), namely, upper, medium and lower. This stimulates us to further examine the physicochemical properties of these three amino acids: (1) Molecular weight: Arg takes the upper (174.20), Leu the medium (131.17) and Ser the lower (105.09). (2) Hydrophobicity: Arg takes the lower as hydrophilic (−4.5), Leu the upper as hydrophobic (3.8), and Ser is medium so that it is neutral (−0.8); this holds for different hydrophobicity scales (*43-45*). (3) Surface area: Arg is the upper (225), Leu the medium (170), and Ser the lower (115). (4) Structure: Leu is α-helix and Ser is turn, whereas Arg is versatile, either α-helix, β-sheet or turn. Consistent with our expectations, these three amino acids have diverse physicochemical properties, presumably responsible for different scenarios to balance physicochemical properties.

Conclusion

In this study, we reorder the four nucleotides according to their emergence in evolution, and apply the organizational rules to devising an algebraic representation for the canonical genetic code. Under a framework of the devised code, we quantify codon and amino acid usages from a large collection of 917 prokaryotic genome sequences, and associate the usages with its intrinsic structure and classification schemes as well as amino acid physicochemical properties. Our results show that the algebraic representation of the code is structurally equivalent to a content-centric organization of the code and that codon and amino acid usages under different classification schemes were correlated closely with GC content, implying a set of rules governing composition dynamics across a wide variety of prokaryotic genome sequences. These results also indicate that codons and amino acids are not randomly allocated in the code, where the six-fold degenerate codons and their amino acids have important balancing roles for error minimization. Therefore, the content-centric genetic code is of great usefulness in deciphering its hitherto unknown regularities as well as the dynamics of nucleotide, codon, and amino acid compositions.

Materials and Methods

Data collection

We retrieved prokaryotic genome sequences from NCBI at ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/. In order to ensure a sufficient sample size for obtaining codon and amino acid frequencies, we excluded species with less than 64 coding sequences. Species with alternative genetic codes were also eliminated from this study. As a consequence, we obtained a total of 917 prokaryotic genomes (including 101 archaea and 806 bacteria). Similar results were obtained after removing the archaea sequences (data not shown). Codon usage presented in this study was computed after eliminating stop codons. The information of these genome sequences as well as codon and amino acid usages is listed in Table S1.

Linear regression analysis

The linear regression uses the least squares approach, implemented by a statistical software package named PAST (*46*). Several relevant statistics are estimated, including slope, intercept, correlation coefficient (R) and two-tailed *P*-value. In our study, the squared correlation coefficient (R^2, or the coefficient of determination) is used, in that R^2 defines the proportion of variance in common between two variables.

Acknowledgements

We thank anonymous reviewers for their critical comments and constructive suggestions on this manuscript. This work was supported by a faculty fund from King Abdullah University of Science and Technology (http://www.kaust.edu.sa) awarded to JY.

Authors' contributions

ZZ collected the datasets, conducted data analyses, and drafted the manuscript. JY supervised the study and revised the manuscript. Both authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1 Crick, F.H. 1968. The origin of the genetic code. *J. Mol. Biol.* 38: 367-379.

2 Jungck, J.R. 1978. The genetic code as a periodic table. *J. Mol. Evol.* 11: 211-224.

3 Barricelli, N.A. 1979. On the origin and evolution of the genetic code. II. Origin of the genetic code as a primordial collector language. The pairing-release hypothesis. *Biosystems* 11: 19-28.

4 Hartman, H. 1995. Speculations on the origin of the genetic code. *J. Mol. Evol.* 40: 541-544.

5 Beland, P. and Allen, T.F. 1994. The origin and evolution of the genetic code. *J. Theor. Biol.* 170: 359-365.

6 Knight, R.D. and Landweber, L.F. 2000. The early evolution of the genetic code. *Cell* 101: 569-572.

7 Freeland, S.J., *et al.* 2000. Early fixation of an optimal genetic code. *Mol. Biol. Evol.* 17: 511-518.

8 Hasegawa, M. and Miyata, T. 1980. On the antisymmetry of the amino acid code table. *Orig. Life* 10: 265-270.

9 Portelli, C., *et al.* 1986. Symmetrical-asymmetrical codons and hydrophobic-hydrophilic amino acids. *Physiologie* 23: 139-143.

10 Copley, S.D., *et al.* 2005. A mechanism for the association of amino acids with their codons and the origin of the genetic code. *Proc. Natl. Acad. Sci. USA* 102: 4442-4447.

11 Sjostrom, M. and Wold, S. 1985. A multivariate study of the relationship between the genetic code and the physical-chemical properties of amino acids. *J. Mol. Evol.* 22: 272-277.

12 Koonin, E.V. and Novozhilov, A.S. 2009. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* 61: 99-111.

13 Jimenez-Montano, M.A. 1999. Protein evolution drives the evolution of the genetic code and *vice versa*. *Biosystems* 54: 47-64.

14 Di Giulio, M. 2000. Genetic code origin and the strength of natural selection. *J. Theor. Biol.* 205: 659-661.

15 Freeland, S.J. and Hurst, L.D. 1998. The genetic code is one in a million. *J. Mol. Evol.* 47: 238-248.

16 Wilhelm, T. and Nikolajewa, S. 2004. A new classification scheme of the genetic code. *J. Mol. Evol.* 59: 598-605.

17 Yu, J. 2007. A content-centric organization of the genetic code. *Genomics Proteomics Bioinformatics* 5: 1-6.

18 Jimenez-Montano, M.A. 2009. The fourfold way of the genetic code. *Biosystems* 98: 105-114.

19 Wu, H.L., *et al.* 2005. Evolution of the genetic triplet code via two types of doublet codons. *J. Mol. Evol.* 61: 54-64.

20 Patel, A. 2005. The triplet genetic code had a doublet predecessor. *J. Theor. Biol.* 233: 527-532.

21 Jukes, T.H. 1973. Possibilities for the evolution of the genetic code from a preceding form. *Nature* 246: 22-26.

22 Hayes, B. 1998. The invention of the genetic code. *Am. Sci.* 86: 8-14.

23 Wong, J.T. 1975. A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci. USA* 72: 1909-1912.

24 Levy, M. and Miller, S.L. 1998. The stability of the RNA bases: implications for the origin of life. *Proc. Natl. Acad.*

Sci. USA 95: 7933-7938.

25 Shapiro, R. 1999. Prebiotic cytosine synthesis: a critical analysis and implications for the origin of life. *Proc. Natl. Acad. Sci. USA* 96: 4396-4401.

26 Reader, J.S. and Joyce, G.F. 2002. A ribozyme composed of only two different nucleotides. *Nature* 420: 841-844.

27 Jimenez-Sanchez, A. 1995. On the origin and evolution of the genetic code. *J. Mol. Evol.* 41: 712-716.

28 Xiao, J.F. and Yu, J. 2007. A scenario on the stepwise evolution of the genetic code. *Genomics Proteomics Bioinformatics* 5: 143-151.

29 Shishido, K. 1988. An occurrence of a noticeable alternating pyrimidine-purine run in the replication origins of tetracycline-resistance plasmids pNSI and pT181. *Nucleic Acids Res.* 16: 1640.

30 Biro, J.C., *et al.* 2003. A common periodic table of codons and amino acids. *Biochem. Biophys. Res. Commun.* 306: 408-415.

31 Bashford, J.D. and Jarvis, P.D. 2000. The genetic code as a periodic table: algebraic aspects. *Biosystems* 57: 147-161.

32 Sanchez, R. and Grau, R. 2006. A novel algebraic structure of the genetic code over the galois field of four DNA bases. *Acta Biotheor.* 54: 27-42.

33 Sanchez, R. and Grau, R. 2009. An algebraic hypothesis about the primeval genetic code architecture. *Math Biosci.* 221: 60-76.

34 Duplij, D. and Duplij, S. 2001. Determinative degree and nucleotide content of DNA strands. *Biophysical Bull. Kharkov Univ.* 525: 86-92.

35 Chargaff, E. 1950. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia* 6: 201-209.

36 Chargaff, E. 1951. Some recent studies on the composition and structure of nucleic acids. *J. Cell Physiol. Suppl.* 38: 41-59.

37 Knight, R.D., *et al.* 2001. 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: 10.

38 Hu, J., *et al.* 2007. Compositional dynamics of guanine and cytosine content in prokaryotic genomes. *Res. Microbiol.* 158: 363-370.

39 Antezana, M.A. and Kreitman, M. 1999. The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. *J. Mol. Evol.* 49: 36-43.

40 Taylor, F.J. and Coates, D. 1989. The code within the codons. *Biosystems* 22: 177-187.

41 Chiusano, M.L., *et al.* 2000. Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. *Gene* 261: 63-69.

42 Haig, D. and Hurst, L.D. 1991. A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* 33: 412-417.

43 Kyte, J. and Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157: 105-132.

44 Wimley, W.C. and White, S.H. 1996. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.* 3: 842-848.

45 Hessa, T., *et al.* 2005. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* 433: 377-381.

46 Hammer, Ø., *et al.* 2001. PAST: paleontological statistics software package for education and data analysis. *Palaeontol. Electronica* 4: 9-17.

Supplementary Material

Table S1

DOI: 10.1016/S1672-0229(11)60004-1