A Content-Centric Organization of the Genetic Code

Jun Yu*

Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China.

The codon table for the canonical genetic code can be rearranged in such a way that the code is divided into four quarters and two halves according to the variability of their GC and purine contents, respectively. For prokaryotic genomes, when the genomic GC content increases, their amino acid contents tend to be restricted to the GC-rich quarter and the purine-content insensitive half, where all codons are fourfold degenerate and relatively mutation-tolerant. Conversely, when the genomic GC content decreases, most of the codons retract to the AU-rich quarter and the purine-content sensitive half; most of the codons not only remain encoding physicochemically diversified amino acids but also vary when transversion (between purine and pyrimidine) happens. Amino acids with sixfold-degenerate codons are distributed into all four quarters and across the two halves; their fourfold-degenerate codons are all partitioned into the purine-insensitive half in favorite of robustness against mutations. The features manifested in the rearranged codon table explain most of the intrinsic relationship between protein coding sequences (the informational content) and amino acid compositions (the functional content). The renovated codon table is useful in predicting abundant amino acids and positioning the amino acids with related or distinct physicochemical properties.

Key words: genetic code, codon, GC content, purine content

The universal table for the canonical genetic code has not been changed much since it was discovered and tabulated (*1–3*), albeit some fancy displays (such as in concentric circles) and mythicized code arrangement (such as the Chinese Eight Diagrams and the binary code). Since the genetic code unites the set of four deoxyribonucleotides—the building blocks of DNA—with the set of twenty amino acids—the primary building blocks of proteins, it serves as an "interpreter" that translates an informational content, a chain of the four-letter code, into basically a physicochemical content composed of a twenty-letter "vocabulary"—a functional content—that vitalizes the "language" and "story" of life on earth.

How does the informational content relate to the functional content? It points to, of course, roles for a set of messenger RNAs (mRNAs) and a set of transfer RNAs (tRNAs), as well as the translational machinery, which is another complex product of evolutionary process, as high school students should know in their biology classes. However, the true decisive matrix or mechanism in a conceptual sense is the genetic code (or the intrinsic relatedness of the codons), regardless when and how the current relationship between codon (encoded by mRNA) and anticodon (decoded by specific tRNA) comes to be. In my opinion, this relationship had been predetermined since the early phase of biogenesis (or the origin of life in a broader sense) in terms of which codons correspond to which amino acid, perhaps through an evolutionary mechanism—interplay of mutations and selections at an individual organism and its population level. In other words, if we reshuffle the codon-anticodon relationship now, it would come to the same result over some evolutionary time scales; it is the decision of the genetic code (or codons) and its encoded amino acids (codon-to-AA) instead of the vehicles (mRNAs and tRNAs). A more precise statement should be: it is the decision of nature in choosing an organism that selected the current codon-to-AA relationship; this lucky organism and its offspring, and maybe only its offspring, became the ancestor of all life on earth. The faithful direct progenies are yet for us to find, but the slightly unfaithful ones are now everywhere and in immense forms on the Planet Earth. An archetypical genetic code might have existed and its relics may have been found, but the canonical genetic code ap-

***Corresponding author.**
E-mail: junyu@genomics.org.cn

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

pears universal since the dawn of life (*4*).

The logic for the renovation

It must be reasons significant enough for me to argue for such a fundamental renovation (Table 1). Let me first very briefly review the minimal functionality of the DNA code. For simplicity, I use A, U, G, and C for adenosine, uridine, guanine, and cytosine, respectively. The code of a DNA sequence for producing proteins has only four basic variables: length, order, GC and AG (purine) contents (which are equivalent to what measured for AU and UC contents), if we put aside the variability of nucleotide sequences over time for the moment. Only the last two variables are relevant to the codon table of the canonical genetic code (hereby referred to as the Table). Unfortunately, the popular Table is assembled merely for a concise and neat manifestation of the codon-to-AA context, and scrabbles clear messages that implicate physicochemical diversity of the amino acids and mutation sustainability of the DNA-borne code. Consequently, there is no reason to turn away new synthesis when the Table is not really in its legitimate form and does not give correct illustrations even when similar displays may have been proposed for different reasoning (*5*). Second, the new arrangement attempts to demonstrate the relationship between the informational and functional contents, concerning variability of GC and AG contents, thus more comprehensive and meaningful as we will see. Third, though a minor point, it is also simpler to be understood and memorized since it makes no effort to distinguish the two purines and the two pyrimidines for the third codon position (cp3). I hope everyone who reads this article is able to memorize the renovated Table precisely for his or her life time.

The code in the new Table is divided into four quarters according to their sensitivity to GC content variations

The criteria in favorite of the new arrangement are twofold. First, I considered only two basic variables, GC and AG contents, although there have been many other hypotheses proposed over the past half century, such as resistance to frameshift errors in translation [*6*, *7*; also see a recent review of the history (*8*)]. It has been too much expectation for the codon-to-AA relationship analyses that have led to numerous hypotheses and debates (*2*, *9*, *10*; the relevant publications are so vast in number that I feel guilty to quote only a few. However, this article is not written as a comprehensive overview of the field but merely as an illustration of the alternative codon table). I do not have a slight doubt that the relatedness in the metabolism of amino acids and codon assignment does exist, but it is the genetic code that unites them in a unique way and brings them together to compose functional contents (encoding proteins). Second, I

Table 1 The renovated table of the genetic code*

	A	U	G	C
A	AAR (K) AAY (N)	UAR (St) UAY (Y)	GAR (E) GAY (D)	CAR (Q) CAY (H)
U	AUR (I,M/Sr) AUY (I)	UUR (L) UUY (F)	GUN (V)	CUN (L)
G	AGR (R) AGY (S)	UGR (St,W) UGY (C)	GGN (G)	CGN (R)
C	ACN (T)	UCN (S)	GCN (A)	CCN (P)

*The codons and their corresponding amino acids are arranged in such a way that extra labels are removed. R and Y stand for the two purines and the two pyrimidines, respectively. When R encodes an amino acid and a nonsense codon (stop, St; start, Sr) within the same quadruplet, A corresponds to the first codon in the parentheses and G for the next. Nucleotides for the first and the second codon positions are labeled at the top and on the left-hand side, respectively. The single letter code for amino acids is used.

cannot agree more with early discoveries in nucleoside chemistry that A and T may be more ancient than G and C because C is too unstable to play a role in the origin of life (*11*, *12*), especially when C may not be essential for the constitution of early functional macromolecules (*13*). Perhaps a transient role in the RNA world for C is more than enough for it to deserve a place in the basic structure of DNA; or in a positive notion, its pairing with G makes DNA more stable than the alternatives.

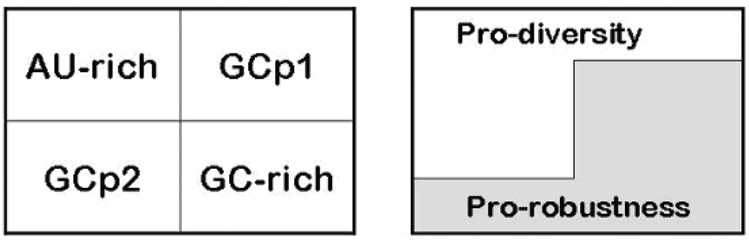
Nevertheless, this particular display divides the Table into four parts; each has its informational and biological sanities: the AU-rich, the GC-rich, and the two intermediate-GC quarters (Table 2A) that are actually not the same regarding to their first and second codon positions (cp1 and cp2). For the sake of convenience, let me call them the GCp1 and GCp2 quarters, referring to the difference of their GC contents at the codon position; codons in the former have G or C at cp1 whereas those in the latter have G or C at cp2. The content-sensitivity of the AU-rich and GC-rich quarters is very obvious if we ignore the third codon position (*14*, *15*). For instance, when the genomic GC content (gGC content) of an organism (say a bacterium) goes up, the former becomes underrepresented and the latter goes overrepresented because every A or T is subjected to be flipped over into G or C. Nature then selects the individual who has made the right choice or eliminates the ones that made the wrong nucleotide flipping before the moment of truth under a given circumstance. The GCp1 and GCp2 quarters remain, by and large, neutral when gGC content varies in a statistic sense overall. One exception is UGG for tryptophane. Although it remains in the same column as other two aromatic amino acids, tyrosine and phenylalanine, it is difficult to know how much it compensates the pressure on UAY and UUY codon pairs when gGC increases, since A or U at cp2 and Y at cp3 have to change simultaneously to turn UAY and UUY into UGG. Certainly, the conversion of UAY to CAY encoding another aromatic amino acid histidine in the GCp1 quarter is possible under pro-GC pressure. In addition, it is also predictable that GC2 (the GC content of cp2) is greater than GC1 (the GC content of cp1) since one of the codons in the GCp2 quarter is a stop codon, despite the fact that general codon usage biases may complicate the real statistics.

The AU-rich quarter possesses the most diversified set of amino acids in terms of their physicochemical properties, containing sixteen codons that encode for seven amino acids as well as two stop and one start signals. In comparison, the GCp1 and GCp2 quarters both encode six amino acids, whereas the GC-rich quarter encodes only four amino acids. The AU-rich quarter is the only quarter that possesses codons for both start and stop signals, leaving us wondering if this quarter might represent the core diversity of the amino acids in building primordial proteins for the early life forms on earth.

The renovated Table divides the code into two halves according to their sensitivity to AG content variations

In addition to the legitimate concern about GC content variations, we can also see a division of the Table into two halves (Table 2B) according to the codon-AA variability between purine and pyrimidine (nucleotide transversions between R and Y) at cp3; it also exhibits a clear separation of the amino acids with fourfold-degenerate codes from those with twofold-degenerate

Table 2 Content-sensitive divisions of the genetic code



A **B**

A. The Table can be divided into four quarters; each has its unique sensitivity to genomic GC content changes: AU-rich, GC-rich, GCp1, and GCp2, and the first two quarters are more sensitive to genomic GC content changes. **B**. The Table can also be divided into two halves: the pro-diversity (transversion-sensitive) half and the pro-robustness (content-insensitive) half with regards to the third codon position (cp3) in particular.

ones albeit the existence of two exceptions, AUR and UGR. This separation further divides the two GC-insensitive quarters (the GCp1 and GCp2 quarters) into two portions. I would like to call these two halves as pro-diversity and pro-robustness (hereby referred to as the PD half and the PR half) based on their functional indications. The clean division allows us to make several observations and predictions. First, three amino acids (serine, arginine, and leucine) that have six codons are partitioned precisely into each of the two halves although they are distributed among all quarters in a seemingly disordered way. This explicit distribution is obvious for the purpose of balancing the three amino acids, which are among the top abundance class by simple statistics such as the average or relative codon usage. An alternative way of explaining this distribution is to assume that these three amino acids were actually selected for the balance due to their roles in maintaining special physicochemical properties (such as catalytic residues) and unique functional domains (such as leucine zippers for transcription factors and the serine-arginine-rich domain for RNA-binding proteins) of proteins. The balancing route involves all four quarters: (1) between the AU-rich and GCp1 quarters (leucine); (2) between the GC-rich and GCp2 quarters (arginine); and (3) within the GCp2 quarter (serine) but across the sub-divided PD and PR halves (serine). As a result of this arrangement, the effect of GC-content increase is reduced through codon conversion rather than amino acid changes. Second, all the nonsense codons appear limited to the PD half. This distribution suggests that the three stop codons (UAA, UAG, and UGA) are readily converted to the corresponding amino acids when GC content goes high, a potential to extend the length of encoded proteins at the 3′-end. Third, the two basic amino acids, arginine and lysine, appear robust against GC-content changes; not only these two amino acids are partitioned sturdily into the PR and PD halves, but also the six arginine codons are divided into the GC-rich quarter and the GCp2 quarter between the two halves. As a contrast, the two acidic amino acids, glutamic acid and aspartic acid, appear taking a different approach—they stay in the GCp1 quarter that is not sensitive to GC-content changes. It is also predictable that these two amino acids must be abundant in the proteins possessing them due to their neutrality against GC-content increase as well as their similarity in chemistry (acidic or negatively charged) and their positions in the Table—they are the most obvious pair resembling a fourfold-degenerate code when charges arose as the only concern in a polypeptide with complex structural constraints (I intend to classify this quadruplet and isoleucine as pseudo-quadruplet). Another likely candidate for achieving high abundance is valine; it has many neighboring amino acids (positioned in the Table with perceptible rationales) that are either similar in hydrophobicity or equivalent in structural characteristics (such as physical dimensions). Finally, staring at the Table, one can easily understand why proline and its codons are sitting at the corner of the GC-rich quarter, and it may only be seriously called upon when GC content goes extremely high.

The Table prioritizes the code to reduce mutation pressure for protein-coding sequences and to maintain functional diversity for proteins as GC and purine contents vary

The renovated Table reveals that the codon arrangement is prioritized to reduce the impact of GC-content variations that fluctuate from 20% to 80% in eubacterial genomes where over 80% of the sequences are protein-coding. In other words, it seizes GC-content variations as the primary parameter. First, it divides the code into two portions, either sensitive (the AU-rich and GC-rich quarters) or insensitive (the GCp1 and GCp2 quarters) to GC-content variations. Second, it confines the high-GC codons as fourfold degenerate to further release the pressure from GC content increases at cp3. Third, it keeps the physicochemically diversified amino acids in the AU-rich quarter (pro-diversity) but leaves the amino acids of the GC-rich quarter to endure mutation pressure (pro-robustness) since they are less likely to be involved in catalytic activities as well as initiation and termination signals (with the exception of the amino acids with sixfold-degenerate codons, especially arginine and serine). Fourth, the function of the GC-insensitive quarters is to protect (or generate in a sense for the origin of the genetic code) the majority of the abundant amino acids in addition to isoleucine in the AU-rich and alanine in the GC-rich quarters. There are certainly more to speculate along this line.

The AG content is the second to be concerned by the Table since it fluctuates only 10% above or below the 50% mark among eubacterial genomes according to the Chargaff's Rule (*16*, *17*). It further divides the GCp1 and GCp2 quarters into purine-variation sensitive and insensitive divisions or the entire Table into two halves. This division draws a clear line that sep-

arates the fourfold-degenerate code with the rest. In the low-diversity (referring to physicochemical properties of amino acids) or the high-robustness (referring to mutation tolerance) half, there are only five amino acids unique to it; each has its subtleties in physicochemical characteristics unique to itself or shared with others. For instance, threonine shares the property of hydroxyl group with serine yet has a slightly extended hydrocarbonic chain. Another example is valine (GUN); it should be one of the most abundant amino acids since it is almost the most flexible amino acid of the quadruplet group, which is capable of replacing leucine (UUR), methionine (AUG), isoleucine (AUY), and even phenylalanine (UUY) when GC content increases and if the mutation-altered protein backbone is embraced by hydrophobicity only.

The third most striking feature in the Table is the clustering of small amino acids aside from the relevance of GC and AG contents. There are several simple measures for the size or volume of the amino acids, such as residue volume (RV, Å^3) (*18*) and accessible surface area (ASA, Å^2) that was calculated for the residue X in the tripeptide G-X-G (*19*). If we rank four smallest amino acids according to their size parameters, they are glycine (RV 60.1 and ASA 75), alanine (RV 88.6 and ASA 115), serine (RV 89.0 and ASA 115), and cysteine (RV 108.5 and ASA 135). The rest of the amino acids are far larger than these four. The next in line is disputable, either aspartic acid (RV 111.1 and ASA 150) or threonine (RV 116.1 and ASA 140), depending on which measurement is preferred. Clearly the most exchangeable pair in size is serine to alanine or *vice versa* when GC content varies.

The ultimate goal of the genetic code is to balance amino acid diversity and robustness to sustain DNA mutation

One essential feature of the Table or the organization of the genetic code is its balancing power. Although the Table divides GC/AG sensitivity vs. insensitivity, amino acid diversity vs. simplicity, and mutation sensitivity vs. tolerance, it seems not favoring one over another. It is predictable that the balance may be severely distorted at least under certain conditions, such as when GC content goes to extremities. The purine content of eubacterial genomes can also go beyond the Chargaff's Rule (*14*), which puts pressure on protein sequence alterations. However, as the Table indicates in its AG-sensitive half, some of the members in this half are there to play relief roles, too. For instance, aspartic acid and glutamic acid are in the same quadruplet, and when a negative charge is essential but not what the size or volume is, a purine to pyrimidine shift in cp3 becomes harmless. To a lesser extent, there are several similar cases in the PD half, including Q/H (size), M/I (hydrophobicity), L/F (hydrophobicity), R/S (polar), W/C (polar), and K/N (polar). This is not farfetched since there has not been a case found possessing a mixed-up feature of hydrophobic vs. hydrophilic or polar vs. non-polar amino acids in the same quadruplet. Sometimes, the obvious seems easier to be overlooked than the obscure.

To sum up, in my thirty years' or so scientific career, I have yet to find another topic that is so fundamental but so misunderstood as the genetic code (even ignoring discussion on the origin of the genetic code). It is critical for biology students to appreciate this rearrangement, to be able to memorize the distribution of the codons (amino acids) in the Table, and to understand the functional indications of the codon positions and their relatedness deeply in order to avoid wasting time to read meaningless publications that have been trying to mystify the genetic code albeit mostly deemed unintentional.

Acknowledgements

The author thanks Dr. Gane K.-S. Wong for long-term collaboration and many stimulating discussions over time related to this topic, and Ms. Wei Gong and Yuanyuan Zhou, Mrs. Zhang Zhang and Kaifu Chen for their assistance in reading and editing the manuscript.

References

1. Nirenberg, M.W. and Matthaei, J.H. 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. USA* 47: 1588-1602.
2. Crick, F.H. 1968. The origin of the genetic code. *J. Mol. Biol.* 38: 367-379.
3. Nirenberg, M. 2004. Historical review: Deciphering the genetic code—a personal account. *Trends Biochem. Sci.* 29: 46-54.
4. Freeland, S.J., *et al.* 2000. Early fixation of an optimal genetic code. *Mol. Biol. Evol.* 17: 511-518.
5. Wilhelm, T. and Nikolajewa, S. 2004. A new classification scheme of the genetic code. *J. Mol. Evol.*

59: 598-605.

6. Woese, C.R. 1965. Order in the genetic code. *Proc. Natl. Acad. Sci. USA* 54: 71-75.
7. Woese, C.R., *et al.* 1966. On the fundamental nature and evolution of the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* 31: 723-736.
8. Bollenbach, T., *et al.* 2007. Evolution and multilevel optimization of the genetic code. *Genome Res.* 17: 401-404.
9. Wong, J.T. 1975. A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci. USA* 72: 1909-1912.
10. Ronneberg, T.A., *et al.* 2000. Testing a biosynthetic theory of the genetic code: fact or artifact? *Proc. Natl. Acad. Sci. USA* 97: 13690-13695.
11. 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.
12. Shapiro, R. 1999. Prebiotic cytosine synthesis: a critical analysis and implications for the origin of life. *Proc. Natl. Acad. Sci. USA* 96: 4396-4401.
13. Reader, J.S. and Joyce, G.F. 2002. A ribozyme composed of only two different nucleotides. *Nature* 420: 841-844.
14. Hu, J.F., *et al.* 2007. Compositional dynamics of guanine and cytosine content in prokaryotic genomes. *Res. Microbiol.* In press.
15. Gu, X., *et al.* 1998. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica* 102-103: 383-391.
16. Chargaff, E. 1951. Structure and function of nucleic acids as cell constituents. *Fed. Proc.* 10: 654-659.
17. Chargaff, E. 1979. How genetics got a chemical education. *Ann. N. Y. Acad. Sci.* 325: 344-360.
18. Chothia, C. 1975. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* 105: 1-12.
19. Zamyatnin, A.A. 1972. Protein volume in solution. *Prog. Biophys. Mol. Biol.* 24: 107-123.