SACRED BOVINES

WAS GENETIC CODING A FROZEN ACCIDENT?

JOHN R. JUNGCK, DOUGLAS ALLCHIN

Genetic coding – whereby DNA sequences specify "information" that can guide the sequencing of amino acids in protein chains – is, quite simply, awe-inspiring. Layers of causality. Visualize, first, how the complexity of human development and cellular activity (from muscular contraction to neural net memory) arises from a mere 25,000 or so proteins, each one functionally based on a unique sculptural shape. Then consider that the diversity of those molecular geographies is based on just a handful – roughly two dozen – basic components: the amino acids. Furthermore, the specification and sequencing of those components are correlated with the arrangement of just four basic units, the familiar nucleotides of DNA. Amazing! Perhaps, just as amazing as an evolved primate writing several dozen Shakespearean plays with just a few dozen alphabetic shapes?

The complexity is formidable. However, at the same time, the arrangement seems extraordinarily fragile. What if the rules for mapping one molecular form to another changed? Any slight modification in the code would apparently lead to ruin. All the messaging would go awry. Life processes would fall apart.

Indeed, this view led some early researchers in the 1960s to view genetic coding as a "frozen accident" in the history of life on Earth (e.g., Crick, 1968, pp. 369, 375). They asserted that once a coding system was in place, its precariousness should prevent it from evolving. "A genetic code, once established, would therefore remain invariant." The "primitive machinery," however arbitrary, would become universal (Hinegardner & Engelberg, 1963, p. 1083). "No new amino acid could be introduced without disrupting too many proteins" (Crick, 1968, p. 375). Indeed, as the code was investigated across many organisms in the 1960s, it seemed they all shared the same codon–amino acid pairings. From bacteria to humans to the viruses of tobacco plants, the genetic code seemed universal. That was hailed as further evidence that all life shared a common ancestor. Historically, this explanation of a frozen accident, along with the view of a necessarily universal genetic code, became "sacred" for many biologists – this month's Sacred Bovine (see also Koonin, 2017; Ribas de Pouplana, et al., 2017; Soll & Bhandary, 2006).

○ Frozen or Variable?

That assumption began to unravel when, with further work, researchers encountered other genetic codes. Not just other messages. Other *genetic codes*.

It began with modest anomalies in the mid-1960s. Lynn Margulis demonstrated that, contrary to widespread assumptions, chloroplasts contain DNA! That finding was soon echoed for mitochondria. That seemed outrageous. What was DNA doing outside the nucleus of eukaryotes? But even more than that: in humans, the mitochondrial genes were unique, not copies of nuclear DNA. Margulis argued that these double-membrane-bound organelles were once independent cells. At some point, they had taken up residence within other, larger cells. Namely, she had exposed another historical sacred bovine with the theory of *endosymbiosis* (*now*, ironically, common textbook knowledge).

Even more surprisingly, molecular biologists found that the mitochondrion's underlying genetic code was distinctive. A handful of the RNA codons were associated with different amino acids, when compared with the coding system found in the nucleus of the same cells (Barrell, Bankier & Drouin, 1979). Then a different set of variants in the code were found in a group of ciliates (Hanyu, et al., 1986) and then algae (Zihala & Eliáš, 2019). As studies have continued over the decades, more "exceptions" continue to be catalogued. Currently, there are at least 33 known variants; 12 of them are found in the mitochondria of various species (Figure 1). Some codes are shared across large groups, while others exist in a single species. The genetic code is plainly not "universal" (Elzanowski & Ostell, 2019; Wikipedia, 2023). Not frozen. Another Sacred Bovine upset.

Biologists also found cases where the genetic code was not restricted to the 20 well-known amino acids. There are two such non-standard amino acids – pyrrolysine and selenocysteine. They seem "hard-coded" in a few of the 33 variants of the genetic code, with their own particular codon. Even more remarkably, in one species of archaebacterium, there is a novel enzyme (and hence, a completely novel gene) that links a tRNA to the amino acid: pyrrolysine aminoacyl-tRNA synthetase (Krzycki, 2005). The "universal" amino acids have turned out to be not strictly universal, either. The ancient code evolved. It was not frozen. One more Sacred Bovine to set by the wayside.

○ An Accident – or Selection at Work?

The diversity of life is full of exceptions. A Darwinian perspective teaches us not to dwell exclusively on the common patterns. The variants, however few, can hide important lessons. In the case of the variants of the genetic code, they offer an occasion to re-examine how the codes may be shaped by natural selection. Here, the field of bioinformatics proves enormously helpful. Mathematical analysis (described below) reveals patterns that tell us that even the standard genetic code, so widely shared phylogenetically, was perhaps not just an arbitrary accident, but reflects selection.

First, it may be helpful to remember that the genetic code is not a symbolic abstraction. In the classroom, we bandy about letters – A, U, C, G – or abbreviations – arg, val, ser, cys. But both

The American Biology Teacher, Vol. 85, No. 9, pp. 550–553, ISSN 0002-7685, electronic ISSN 1938-4211. DOI: https://doi.org/10.1525/abt.2022.85.9.550.

CGG (Arg -> Leu)	*Atractomorpha echinata*
UCA (Ser -> Stop)	*Kirchneriella aperta*
UCG (Ser -> Stop)	*Monoraphidium neglectum*
	Mychonastes homosphaera
	Ourococcus multisporus
	Pseudomuriella schumacherensis
	Raphidocelis subcapitata
AGG (Arg -> Ala)	*Bracteacoccus aerius*
UCA (Ser -> Stop)	*Bracteacoccus minor*
	Chlorotetraedron incus
AGG (Arg -> Met)	*Chromochloris zofingiensis*
CGG (Arg -> Leu)	
UCA (Ser -> Stop)	
UCG (Ser -> Stop)	
AGA (Arg -> Ala)	*Rotundella rotunda*
AGG (Arg -> Ala)	
UCA (Ser -> Stop)	
AGA (Arg -> Ala)	*Coelastrum* sp. F187
AGG (Arg -> Ala)	*Coelastrella* sp. M60
UAG (Stop -> Leu)	*Coelastrella* sp. UTEX B 3026
UCA (Ser -> Stop)	*Coelastrella* sp. YC001
UCG (Ser -> Stop)	*Graesiella* sp. 549
	Hariotina reticulata
	Hariotina sp. MMOGRB0030F
	Pectinodesmus pectinatus
	Tetradesmus obliquus
	Tetradesmus obliquus KS3-2
	Tetradesmus obliquus UTEX 78
AGG (Arg -> Ala)	*Neochloris aquatica*
UAG (Stop -> Ala)	
UCA (Ser -> Stop)	
UAG (Stop -> Ala)	*Hydrodictyon reticulatum*
UCA (Ser -> Stop)	*Lacunastum gracillimum*
	Pediastrum angulosum ACOI 1354
	Pediastrum duplex EL0201CT
	Pediastrum duplex Group I
	Pediastrum duplex Group II
	Pediastrum duplex SL0404MN
	Pseudopediastrum boryanum
	Pseudopediastrum integrum
	Pseudopediastrum sp. CL0201VA
	Stauridium tetras

Figure 1. Variant genetic codes (Zihala & Eliáš, 2019).

amino acids and nucleotide bases are physical entities, not linguistic abstractions. They have their causal effect in the material world. The genetic code is only a metaphor. Molecular interactions are primarily about shape (sizes and bonds), not symbols.

Second, historically "the" genetic code was "cracked" through a sort of black box approach. Put in a poly-U mRNA strand and out comes a polyphenylalanine chain. Conclusion: in symbolic shorthand, "UUU" yields "Phe." Unfortunately, this approach – and the textbook chart based on it — did not uncover *how the code itself worked.* Too often textbooks focus chiefly on translation (assembling the protein), while the critical step where the nucleotides are deciphered receives little attention. Namely, how do tRNAs bridge the gap of two molecular "languages" – of nucleotide base shapes and amino acid shapes? How does the cell "know" which amino acid to load onto which tRNA? The rest is mechanistic assembly.

This might raise the provocative question, "where exactly in the cell is the genetic code?" Of course, there is no reference table stored in cells. There is no miniature brain "reading" each anti-codon and consciously fetching the corresponding amino acid. Rather, the deciphering is achieved "blindly": by shape and by chemical interactions, not something akin to visual recognition. For each amino acid, there is a uniquely shaped enzyme that couples the tRNA's anti-codon shape (at one location) with the unique features of a corresponding amino acid (at another site): these are the *amino acyl-tRNA synthetases* (or AARSs) described in more advanced texts (Figure 2). The conventional textbook chart is a human construct, just like the periodic table of the elements.

Of course, "errors" may occur. The DNA may mutate. The amino acid may misload onto the tRNA. The mRNA may misalign with a nonmatching tRNA. The AARS itself may evolve and change shape at either recognition site. Each error leads to a distinct change in the coding process. What will be the fate of any variant? Again, a bioinformatics approach can help, by analyzing the structure of the genetic code, the tRNAs, and the physicochemical properties of the amino acids.

What, specifically, is the result of single base mutations in the DNA? Most such mutations in the DNA are conservative.

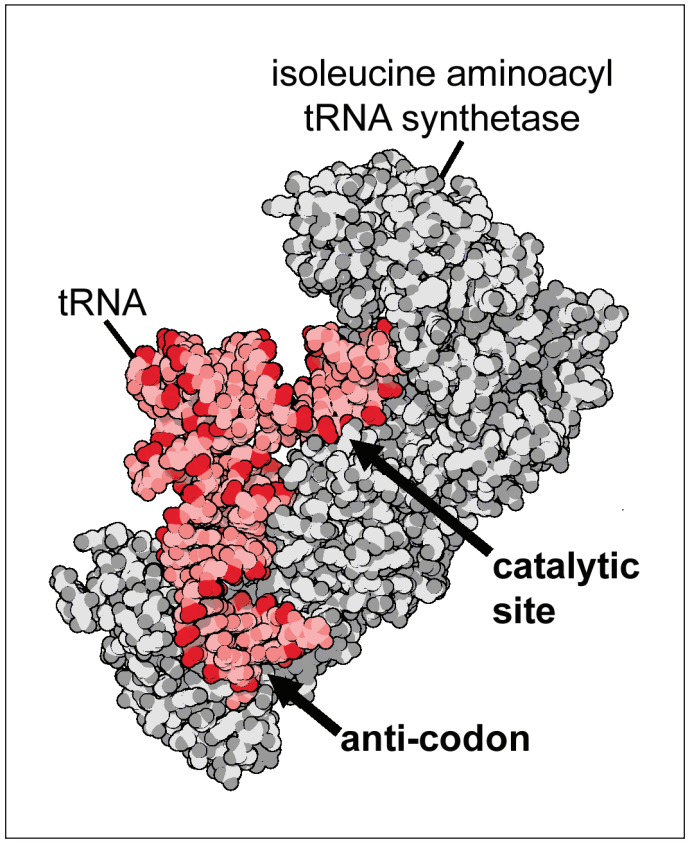

Figure 2. A tRNA molecule embedded in isoleucine aminoacyl-tRNA synthetase. The amino acid is not shown. Image adapted from David S. Goodsell and the RCSB PDB's "Molecule of the Month" (2015).

That is, the changed codon yields a chemically similar or even identical amino acids. Why? Amino acids may be characterized in three fundamental dimensions. Are they large or small? Are they hydrophobic or hydrophilic? Are they interior or exterior? (See Figure 3 for how the 20 common amino acids are sorted by these criteria.)

Based on these categories, we can examine whether a single nucleotide substitution (mutation) results in a significantly different amino acid. The substitutions are not random. The codons are clustered by amino acid type. One finds three major groups: the hydrophobic group (valine, leucine, isoleucine, alanine); the polar group (cysteine, methionine, asparagine, glutamine, serine, threonine); and the charged group (arginine, lysine, histidine, glutamic acid, aspartic acid). Again, evidence of selection at work.

Moreover, the effects of substitutions differ substantially, depending on the position of the mutation. Mutations in the 3rd position of the codon often result in no change in amino acid (Figure 4, last column). In other cases, what matters is merely whether the 3rd base is a purine (A, G) or pyrimidine (C, U). This indicates that, spatially, the 3rd base is not a very strong determinant in how the AARS recognizes or binds with the tRNA molecule. Such mutations are less likely to be selected against. However, a mutation in the 2nd position leads to major changes. Indeed, the 2nd base alone tends to determine the character of the amino acid (Figure 4, 2nd column). U generally codes for a hydrophobic amino acid, and A for a hydrophilic one. C is associated with one of the four small amino acids (and G varies). In the case here, a change in the 1st base, is far less dramatic. It yields a similar amino acid, as is usually the case. Ultimately, the genetic code buffers against adverse mutations, ones that would significantly change the three-dimensional configuration of the resulting protein. The odds that such an error-resistant code would arise exclusively by chance have been calculated as less than 1 in 100,000 (Koonin, 2017). Historical contingency seems to have acted in concert with selection.

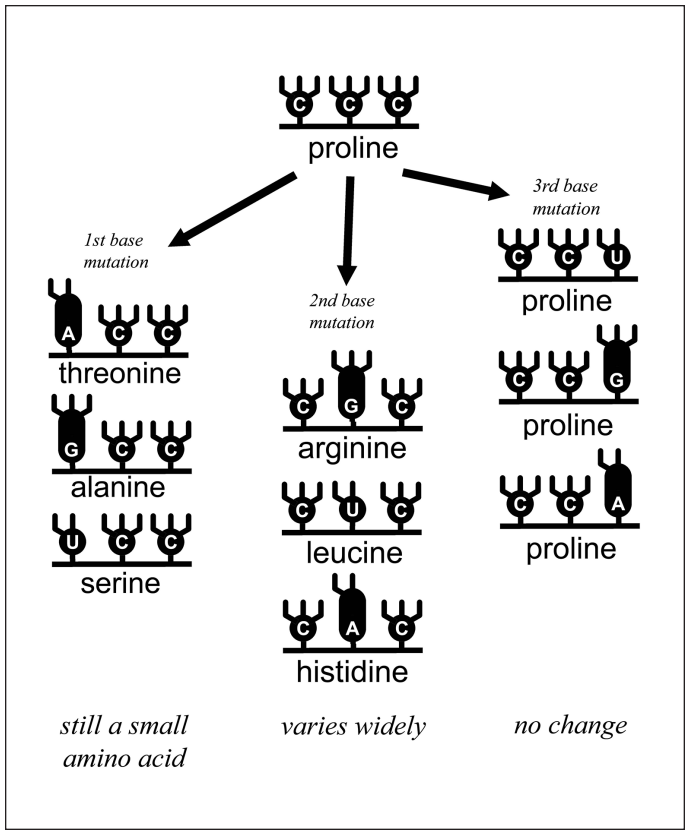

Figure 4. Effect on amino acid of single base mutation for the codon "CCC" (proline). The 3rd base results in no change, indicating that its shape is relatively unimportant in determining which amino acid the AARS binds with. A change in the 2nd base leads to major changes. It is the strongest determinant of the amino acid type. A change in the 1st base leads to changes, but typically of a similar amino acid.

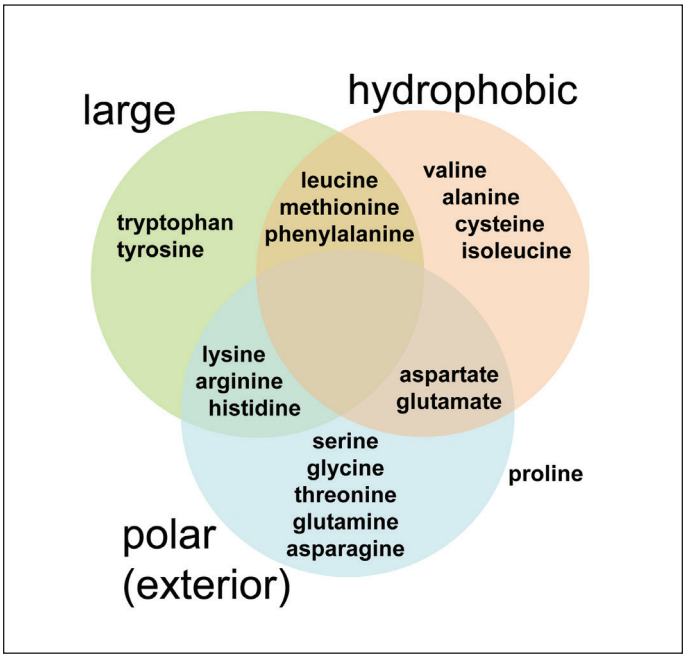

Figure 3. Amino acids sorted by their physicochemical properties.

Second, why do some amino acids have more codons than others? If you have ever played Scrabble®, you will remember that some letters are worth more (Q and Z are worth 10 points), while others are worth less (A, E, I, O, N, T and U are worth only 1 point). These choices in the game design were based on the letters' frequency. The standard genetic code embodies a similar pattern, yielding greater efficiency of protein assembly (Jungck, 2009). Namely, cells use more readily available amino acids, such as serine and leucine, and they have more codons. Diffusion time of tRNAs is reduced. Infrequent amino acids, such as methionine and tryptophan, whose codons are less common, are associated with longer waiting times. The genetic code seems to have evolved based on efficiency. Again, that does not seem wholly arbitrary, or accidental.

Third, why are particular codons associated with particular amino acids? Again, selection seems likely. There seems to be a physicochemical basis linking anticodons in the tRNAs to their particular amino acids, at least based on the first two nucleotides of the anticodon (Jungck, 1978, among others). While the hypothesis remains somewhat controversial, Cockell (2018) argues for "ancient attractions between amino acids and little strands of RNA, perhaps even before transfer RNAs became the intermediary between them, … Those affinities laid the groundwork for the link between the decoding of RNA into proteins."

○ A More Accurate View of Genetic Coding, and of Ourselves

So, contrary to once-widespread assumptions, "the" genetic code was not frozen into a now-universal code. There are *multiple* genetic codes across the evolutionary tree. That is true even within our own cells – another reminder of the endosymbiotic origin of the mitochondria. We are hybrid organisms, with different genetic codes even within our own bodies.

Nor are the codes completely arbitrary. They show unmistakable evidence of natural selection. For example, the middle base of the codon is a strong determinant of the chemical nature of the coded amino acid. This correspondence has remained stable, while other coding elements have changed. In addition, more frequent amino acids have more codons. And the genetic codes are partly degenerate (some bases redundant). Further, the codes seem buffered against deleterious single base mutations. Ultimately, the code(s) have evolved based on efficiency and protection against mutation. The genetic codes do not seem entirely "accidental," as Crick once supposed. They seem to have evolved through natural selection, just like other traits of organisms.

References

Barrell, B. G., Bankier, A. T., & Drouin, J. (1979). A different genetic code in human mitochondria. *Nature*, *282*, 189–194.

Cockell, C. (2018). *The Equations of Life: How Physics Shapes Evolution*. Atlantic Books.

Crick, F. H. (1968). The origin of the genetic code. *Journal of Molecular Biology*, *38*, 367–379.

Elzanowski, A. & Ostell, J. (2019). The genetic codes. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

Goodsell, D. S., Dutta, S., Zardecki, C., Voigt, M., Berman, H. M. & Burley, S. K. (2015). The RCSB PDB 'Molecule of the Month': Inspiring a molecular view of biology. *PLoS Biology*, *13*(5), e1002140. https://doi.org/10.1371/journal.pbio.1002140

Hanyu, N., Kuchino, Y., Nishimura, S., & Beier, H. (1986). Dramatic events in ciliate evolution: Alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two Tetrahymena tRNAsGln. *EMBO Journal*, *5*, 1307–1311.

Hinegardner, R. T., & Engelberg, J. (1963). Rationale for a universal genetic code. *Science*, *142*, 1083–1055.

Jungck, J. R. (1978). The genetic code as a periodic table. *Journal of Molecular Evolution*, *11*, 211–224.

Jungck, J. R. (2009). Genetic codes as codes: Towards a theoretical basis for bioinformatics. In R. Mondani (ed.), *BIOMAT 2008* (pp. 300–337). World Scientific.

Koonin, E. V. (2017). Frozen accident pushing 50: Stereochemistry, expansion, and chance in the evolution of the genetic code. *Life*, 7, 22. https://doi.org/10.3390/life7020022.

Krzycki, J. A. (2005). The direct genetic encoding of pyrrolysine. *Current Opinion in Microbiology*, *8*, 706–712.

Ribas de Pouplana, L., Torres, A. G. & Rafels-Ybern, A. (2017). What froze the genetic code? *Life*, 7(2), 14. https://doi.org/10.3390/life7020014

Soll, D., & Bhandary, U. L. (2006). The genetic code: Thawing the frozen accident. *Journal of Bioscience*, *31*(4), 459–463.

Wikipedia. (2023). List of genetic codes. https://en.wikipedia.org/wiki/List_of_genetic_codes; and https://en.wikipedia.org/wiki/DNA_and_RNA_codon_tables#Alternative_codons.

Zihala, D., & Eliáš, M. (2019). Evolution and unprecedented variants of the mitochondrial genetic code in a lineage of green algae. *Genome Biology and Evolution*, *11*, 2992–3007.

JOHN R. JUNGCK is at the Delaware Biotechnology Institute of the University of Delaware. DOUGLAS ALLCHIN is at the Minnesota Center for the Philosophy of Science at the University of Minnesota.