ODE TO THE CODE

Brian Hayes

The genetic code was cracked 40 years ago, and yet we still don't fully understand it. We know enough to read individual messages, translating from the language of nucleotide bases in DNA or RNA into the language of amino acids in a protein molecule. The RNA language is written in an alphabet of four letters (A, C, G, U), grouped into words three letters long, called triplets or codons. Each of the 64 codons specifies one of 20 amino acids or else serves as a punctuation mark signaling the end of a message. That's all there is to the code. But a nagging question has never been put to rest: Why this particular code, rather than some other? Given 64 codons and 20 amino acids plus a punctuation mark, there are 10^{83} possible genetic codes. What's so special about the one code that—with a few minor variations—rules all life on Planet Earth?

The canonical nonanswer to this question came from Francis Crick, who argued that the code need not be special at all; it could be nothing more than a "frozen accident." The assignment of codons to amino acids might have been subject to reshuffling and refinement in the earliest era of evolution, but further change became impossible because the code was embedded so deeply in the core machinery of life. A mutation that altered the codon table would also alter the structure of every protein molecule, and thus would almost surely be lethal. In other words, the genetic code is the qwerty keyboard of biology—not necessarily the best solution, but too deeply ingrained to be replaced or improved.

There has always been resistance to the frozen-accident theory. Who wants to believe that the key to life is so arbitrary and *ad hoc*? And there is evidence that the accident is not quite frozen. Certain protozoa, bacteria and intracellular organelles employ genetic codes slightly different from the standard one, hinting that changes to codon assignments are not impossible after all. And if the code is subject to change, then it must also be subject to natural selection, which in turn suggests the possibility of ongoing improvement. Perhaps ours is not the very best of all possible codes, but after four billion years of evolution it ought to be a pretty darn good one.

The urge to find something singular and superlative about the code was already evident even before it was deciphered. For several years before experiments began to reveal the true structure of the genetic code, theorists were at liberty to dream up codes of their own. Some of the proposals were so ingenious that the real code seemed a bit disappointing. An earlier column in this series (January–February 1998) described that era of imaginary genetic engineering. But the creative thinking did not end with the publication of the codon table; indeed speculation seems to have been inhibited very little by the constraints of mere fact. This sequel is meant to bring the story up to date, covering both the biological mainstream and a few ideas from wilder shores.

Brian Hayes is Senior Writer for American Scientist. *Address: 211 Dacian Avenue, Durham, NC 27701; bhayes@amsci.org*

Egged on by Error

Early guesses about the nature of the code often started from an assumption that it would maximize information density. One conjecture had each nucleotide base spelling out three messages at once. The concern with efficiency turned out to be misplaced; information density is not a very high priority for most organisms. The concept that has replaced efficiency as the great desideratum in genetic coding is error-tolerance, or robustness. In one way or another, the code is thought to minimize the incidence and the consequences of errors in the transmission of genetic information, so that meaning can be recovered even from garbled messages.

Among the many ways that genetic signals could go awry, two kinds of errors have been singled out for attention: mistranslations and mutations. Errors in translation disrupt the reading of the genetic message—the flow of information from DNA to RNA and then to protein—but they leave the DNA itself intact. Translation errors were probably of great importance early in the history of life, when the machinery of protein synthesis was imprecise. Mistranslations are less frequent now, and less harmful. Each error disables only a single protein molecule. Mutations are another matter: They alter the DNA, the permanent genetic archive. Whereas a translation

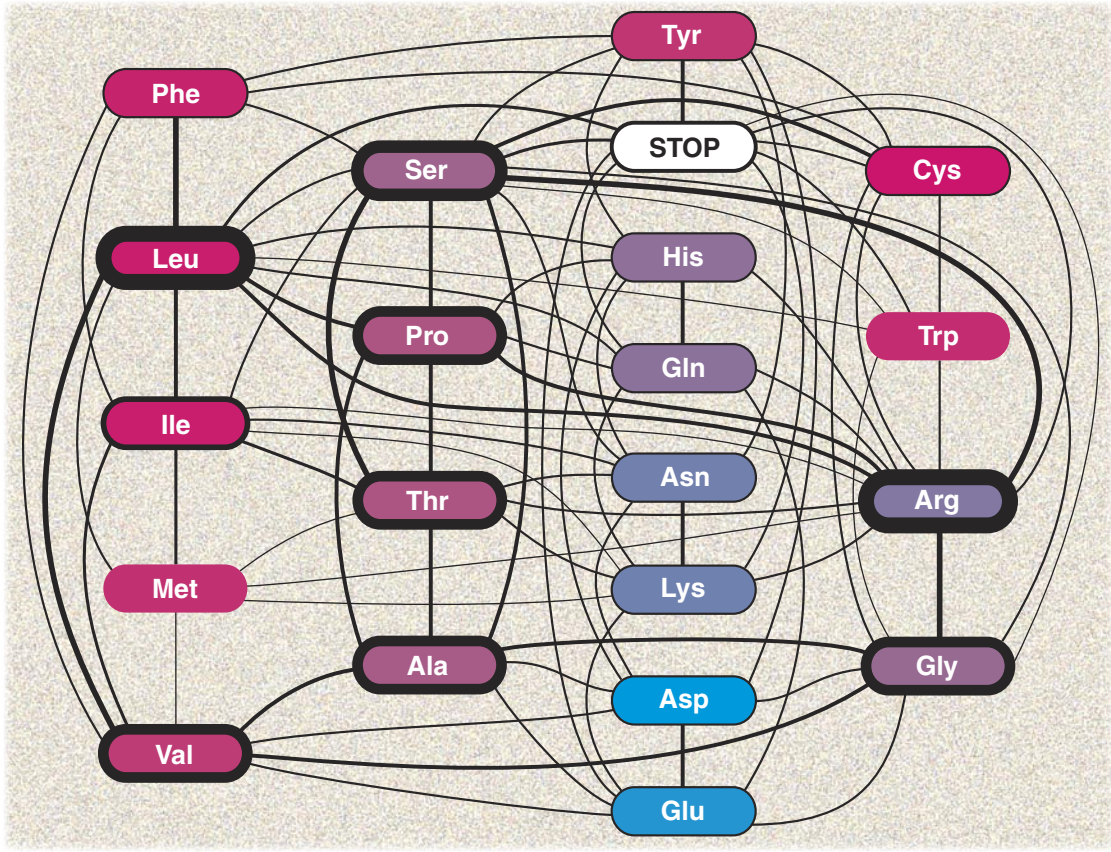

Figure 1. Consequence of mutations and other errors in transmitting genetic information are ameliorated by the structure of the genetic code. The mutations themselves, which take place in the DNA, are not visible in this diagram; the network of nodes and arcs shows the effect of the mutations on the amino acid constituents of proteins. The nodes are the 20 amino acids, plus the "STOP" symbol that marks the end of a gene. Two nodes are connected by an arc if a single mutation can result in the substitution of one amino acid for the other. The thickness of the arc is proportional to the number of distinct mutations that can cause a given substitution; the thickness of the black border surrounding a node indicates the number of mutations that leave the identity of the amino acid unchanged. For example, four mutations convert a DNA "codon" for alanine (Ala) into one for valine (Val), but only two mutations convert alanine into aspartic acid (Asp). Twelve mutations convert one alanine codon into another. The colors of the nodes represent an important property of amino acids: As the colors go from blue to red, the amino acids go from hydrophilic to hydrophobic. The genetic code seems to be organized so that common substitutions cause little change in this property.

error is like an inkblot marring one copy of a book, a mutation is a flaw in the printing plate, reproduced in every copy. The simplest "point" mutations substitute one nucleotide for another at a single site on the DNA (with a corresponding change on the opposite strand).

The idea that fault tolerance might shape the genetic code arose as soon as biologists got their first glimpse of the codon table. The mapping from codons to amino acids is highly degenerate: In many cases multiple codons specify the same amino acid. But the synonymous codons are not just scattered haphazardly across the table; they clump together. Because of these clusters, a misreading or mutation has a better-than-average chance of producing a new codon that still translates into the same amino acid.

Closer examination of the table—with some knowledge of amino acid chemistry—revealed another possible strategy for coping with errors. When a change to a single nucleotide does not yield the same amino acid, it nonetheless has a good chance of producing one with similar properties. For example, all the codons with a middle nucleotide of U correspond to amino acids that are hydrophobic, or water-repellent, a trait governing how the chain of amino acids in a protein molecule folds up in the aqueous environment of the cell. Thus at least two-thirds of the time a point mutation in one of these codons will either leave the identity of the amino acid unchanged or will substitute another hydrophobic amino acid.

Reshuffling the Deck of Codons

As early as 1969, Cynthia Alff-Steinberger of the University of Geneva began trying to quantify the code's resilience to error by means of computer simulation. The basic idea was to randomly generate a series of codes that reshuffle the codon table but retain certain statistical properties, such as the number of codons associated with each amino acid. Then the error-resistance of the codes was evaluated by generating point mutations that caused amino acid substitutions.

A code scored well if the erroneous amino acids were similar to the original ones. With the computing facilities available in the 1960s, Alff-Steinberger was able to test only 200 variant codes. She concluded that the natural code tolerates substitutions better than a typical random code.

A decade later J. Tze-Fei Wong of the University of Toronto approached the same question from another angle—and reached a different conclusion. Instead of generating many random codes, he tried a hand-crafted solution, identifying the best substitution for each amino acid. Wong found that the substitutions generated by the natural code are less than half as close, on average, as the best ones possible. This result was taken as evidence that the code has *not* evolved to maximize error tolerance. But Wong did not attempt to find a complete, self-consistent code would generate all the optimal substitutions.

Returning to studies of random codes, David Haig and Laurence D. Hurst of the University of Oxford generated 10,000 of them in 1991, keeping the same blocks of synonymous codons found in the natural code but permuting the amino acids assigned to them. The result depended strongly on what criterion was chosen to judge the similarity of amino acids. Using a measure called polar requirement, which indicates whether an amino acid is hydrophobic or hydrophilic, the natural code was a stellar performer, better than all but two of the 10,000 random permutations. But in other respects the biological code was only mediocre; 56 percent of the random codes did a better job of matching the electric charge of substituted amino acids.

Focusing on the encouraging result with polar requirement, Hurst and Stephen J. Freeland (now at the University of Maryland Baltimore County) later repeated the experiment with a sample size of 1 million random codes. Using the same evaluation rule as in the smaller simulation, they found that 114 of the million codes gave better substitutions than the natural code when evaluated with respect to polar requirement. Then they refined the model. In the earlier work, all mutations and all mistranslations were considered equally likely, but nature is known to have certain biases—some errors are more frequent than others. When the algorithm was adjusted to account for the biases, the natural code emerged superior to every random permutation with a single exception. They published their results under the title "The genetic code is one in a million."

But still there was the question of whether polar requirement is the right criterion for estimating the similarity of amino acids. Choosing the one factor that gives the best result and ignoring all others is not an experimental protocol that will convince skeptics. This issue was addressed in a further series of experiments by Freeland and Hurst in collaboration with Robin D. Knight and Laura F. Landweber of Princeton University. Rather than try to deduce nature's criteria for comparing amino acids, they inferred it from data on actual mutations. If two amino acids are often found occupying the same position in variant copies of the same protein, then it seems safe to conclude that the amino acids are physiologically compatible. Conversely, amino acids that are never found to occupy the same position would not be likely substitutions in a successful genetic code. There is a circularity to this formulation: The structure of the genetic code helps determine which substitutions are seen most often,

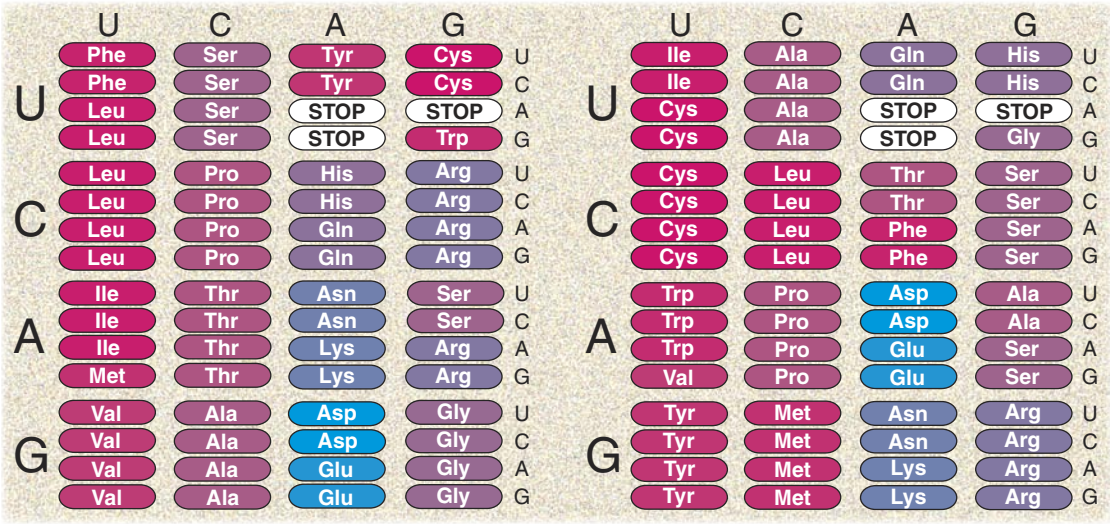

Figure 2. Standard genetic code *(left)* translates from the language of nucleotide bases in DNA or RNA into the amino acid language of proteins. The four bases of RNA are represented by the letters U, C, A, G; each of the 64 possible triplets, or codons, formed from these letters specifies an amino acid. To translate a codon, match the three bases in sequence to the symbols along the left margin, the top and the right margin of the table; for example, AUG encodes methionine (Met). As in Figure 1, colors indicate each amino acid's position along the hydrophilic-hydrophobic axis. The genetic code at right does not exist in nature but emerged in a computer experiment by Stephen J. Freeland and Laurence D. Hurst as "one in a million"—the only artificial code out of a million random trials that performed better than the natural code in making amino acid substitutions that are similarly hydrophilic or hydrophobic.

and then the frequencies of substitutions serve to rank candidate genetic codes. Freeland and his colleagues argue that they can break the cycle by choosing an appropriate subset of the mutation data, including only proteins at substantial evolutionary distance, which should be separated by many mutations.

Using this bootstrap criterion, Freeland and his colleagues compared the biological code with another set of a million random variations. The natural code emerged as the uncontested champion. They wrote of the biological code: "...it appears at or very close to a global optimum for error minimization: the best of all possible codes."

Antievolutionists

The idea that the genetic code is evolving under pressure to ameliorate errors—or indeed that it is evolving at all—has not won universal assent. Some cogent objections were set forth as early as 1967 by Carl R. Woese of the University of Illinois at Urbana-Champaign. Among other points, he noted that if a trait is actively evolving, you would expect to see some variation. In particular he called attention to the various "extremophiles" that live at high temperature, high salt concentration, and so on. These organisms tend to have unusual proteins and unusual nucleic acids, but they all have the standard genetic code.

The few variant codes known in protozoa and organelles are thought to be offshoots of the standard code, but there is no evidence that the changes to the codon table offer any adaptive advantage. In fact, Freeland, Knight, Landweber and Hurst found that the variants are inferior or at best equal to the standard code. It seems hard to account for these facts without retreating at least part of the way back to the frozen-accident theory, conceding that the code was subject to change only in a former age of miracles, which we'll never see again in the modern world.

Another challenge to the error-reduction hypothesis is the difficulty of showing causation in an evolutionary context. Even if the pattern of codon assignments is consistent with such a mechanism, the same pattern might have arisen in some other way.

Computer experiments like Alff-Steinberger's and Freeland's reveal nothing about pathways of evolution. A program churning out a million random genetic codes is not what you expect to see in nature. To simulate the step-by-step process of mutation and selection is much more demanding; after all, the biosphere has been working at it for a few billion years. Nevertheless, models of this kind are being attempted. Guy Sella and David H. Ardell of Stanford University are running a simulation that includes both a nucleic acid genotype and a protein phenotype, linked by a mutable genetic code. They point out that change can be introduced into the genetic code without utterly disrupting cell metabolism if there are multiple codons for a given amino acid,

Figure 3. A dodecahedral symmetry of the genetic code, discovered by Mark White of Bloomington, Indiana, is embodied in White's soccer-ball–like toy. The RNA bases are inscribed on the 12 faces of the dodecahedron. To rad an amino acid assignment from the ball, make a tour of three adjacent bases. The encoded amino acid will be found adjacent to the first base, along the axis leading to the second base, and inside the triangle formed with the third base. For example, in the part of the ball visible here, a tour from A to G to U specifies serine, but going from A to U to G corresponds to methtionine.

and some of them fall into disuse; these rarely used codons are then free to take on new roles. The mechanism is analogous to the gene duplication that often precedes evolutionary divergence of proteins: One copy of the gene carries on the original function, allowing the other to explore new territory. Thus degeneracy or redundancy is not just an accidental feature of the code but is necessary to allow scope for evolution.

Code On, Codon

Solomon W. Golomb of the University of Southern California, who was a central figure in the first round of speculations about the genetic code, has summed up the spirit of that era: The approach taken in those days was to ask, "How would Nature have done it, if she were as clever as I?" Now that we know how nature has done it, you might think that the period of freewheeling conjecture would be over, but I am pleased to report that there is no lack of adventurous ideas about patterns and structures in the genetic code. Here are just a few of the ideas in circulation.

One of the themes of the earlier period was the need to find some compelling relation between the numbers 64 and 20. And this quest had spectacular successes: In at least two schemes, the 64 codons could specify exactly 20 amino acids, neither more nor less. The mathematics was so beautiful, it was hard to believe nature would pass up an opportunity to make use of it. Pierre Béland and T. F. H. Allen of the St. Lawrence National Institute of Ecotoxicology in Montreal argue that nature did *not* miss the opportuni-

ty. They propose a primordial genetic code in which information was read from both strands of the DNA at once, and all messages were palindromic, so that they could be read in either direction. Under these conditions, meaning can be assigned to only 20 of the 64 triplets.

A double-stranded translation system may sound outlandish, and yet there are hints that the "antisense" strand of DNA may be more than just a placeholder. Jaromir Konecny, Michael Schöniger and G. Ludwig Hofacker of the Technical University of Munich point out that a rough symmetry of the genetic code creates a kind of antigene opposite every normal gene. Wherever the sense strand calls for a hydrophilic amino acid, the antisense strand (read in the opposite direction) is likely to code for a hydrophobic one. It's even possible that some of these antisense pseudogenes are transcribed *in vivo*. William F. Pendergraft III and six colleagues at the University of North Carolina at Chapel Hill have recently detected immunological reactions to one such antisense protein.

More generally, there is growing recognition that the genetic code may encompass more information than just the simple mapping from codons to amino acids. Synonymous codons may not always be completely equivalent. It's certainly true that codon frequencies are not random or uniform. Among the several codons that specify a given amino acid, some may be common and some rare, and these usage biases can vary both within and between genomes. The biases probably help to regulate the rate of protein synthesis: If the transfer RNA that matches a codon is rare, then transcription of genes including that codon will be slowed. For some proteins there is evidence that such pace-setting codons help ensure correct folding of the amino acid chain.

Another fertile area is the search for symmetries and patterns in the genetic code. The standard table of codon assignments derives from the obvious representation of the triplet code as a $4\times4\times4$ cube. Several authors, observing that 64 is equal not only to 4^3 but also to 2^6, suggest organizing the codon table as a six-dimensional $(2\times2\times2\times2\times2\times2)$ hypercube. A mutation is a movement from one vertex to an adjacent vertex in this structure. The geometry is intriguing, and there are interesting connections with Gray codes and even with the *I Ching*, but I'm not so sure that biologists will find the concept useful.

Not every interesting idea takes the form of a paper in the *Journal of Theoretical Biology*. Another quite different geometrical interpretation of the genetic code has been presented to the world in the form of a design for a toy. Mark White, a physician and inventor in Bloomington, Indiana, discovered that the genetic code can be represented succinctly on a dodecahedron (a solid whose surface consists of 12 pentagons) or its dual the icosahedron (made up of 20 triangles). Each face of the dodecahedron is labeled with one of the four nucleotides, each of which appears three times. Any grouping of three adjacent faces, read in the right order, generates the appropriate amino acid. White has made prototypes of toys that incorporate this design. He observes that the icosahedral model is closely related to the very first proposal for a triplet genetic code, the "diamond code" devised in 1955 by George Gamow. This neatly closes the circle and takes us back to the beginning of the story.

Bibliography

Alff-Steinberger, C. 1969. The genetic code and error transmission. *Proceedings of the National Academy of Sciences of the U.S.A.* 64:584–591.

Béland, Pierre, and T. F. H. Allen. 1994. The origin and evolution of the genetic code. *Journal of Theoretical Biology* 170:359–365.

Cortazzo, Patricia, Carlos Cerveñansky, Mónica Marín, Claude Reiss, Ricardo Ehrlich and Atilio Deana. 2002. Silent mutations affect in vivo protein folding in *Escherichia coli*. *Biochemical and Biophysical Research Communications* 293:537–541.

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

Freeland, Stephen J., and Laurence D. Hurst. 1998. The genetic code is one in a million. *Journal of Molecular Evolution* 47:238–248.

Freeland, Stephen J., Robin D. Knight, Laura F. Landweber and Laurence D. Hurst. 2000. Early fixation of an optimal genetic code. *Molecular Biology and Evolution* 17(4):511–518.

Freeland, Stephen J., Tao Wu and Nick Keulmann. 2003. The case for an error minimizing standard genetic code. *Origins of Life and Evolution of the Biosphere* 33: 457–477.

Golomb, Solomon W. 1980. Cryptographic reflections on the genetic code. *Cryptologia* 4(1):15–19.

Haig, David, and Laurence D. Hurst. 1991. A quantitative measure of error minimization in the genetic code. *Journal of Molecular Evolution* 33:412–417.

Hayes, Brian. 1998. The invention of the genetic code. *American Scientist* 86:8–14.

Jiménez–Montaño, Miguel A., Carlos R. de la Mora–Basáñez and Thorsten Pöschel. 1995. On the hypercube structure of the genetic code. In *Proceedings of the Third International Conference on Bioinformatics & Genome Research*, ed. Hwa A. Lim and Charles A. Cantor, Singapore: World Scientific, pp. 445–455.

Konecny, Jaromir, Michael Schöniger and G. Ludwig Hofacker. 1995. Complementary coding conforms to the primeval comma-less code. *Journal of Theoretical Biology* 173:263--270.

Morimoto, Susumu. 2002. A periodic table for genetic codes. *Journal of Mathematical Chemistry* 32:159–200.

Pendergraft, William F. III, Gloria A. Preston, Ruchir R. Shah, Alexander Tropsha, Charles W. Carter, Jr., J. Charles Jennette and Ronald J. Falk. 2004. Autoimmunity is triggered by cPR-3(105–201), a protein complementary to human autoantigen proteinase-3. *Nature Medicine* 10:72–79.

Sella, Guy, and David H. Ardell. 2002. The impact of message mutation on the fitness of a genetic code. *Journal of Molecular Evolution* 54:638–651.

White, Mark. Undated. Maximum symmetry in the genetic code: The Rafiki map. Unpublished manuscript. http://www.codefun.com/Images/Genetic/Max/Sym300dpi.pdf

Woese, Carl R. 1967. *The Genetic Code: The Molecular Basis for Genetic Expression*. New York: Harper & Row.

Wong, J. Tze-Fei. 1980. Role of minimization of chemical distances between amino acids in the evolution of the genetic code. *Proceedings of the National Academy of Sciences of the U.S.A.* 77:1083–1086.