THE INVENTION OF THE GENETIC CODE

Brian Hayes

A reprint from

American Scientist

the magazine of Sigma Xi, the Scientific Research Society

Volume 86, Number 1
January–February, 1998
pages 8–14

THE INVENTION OF THE GENETIC CODE

Brian Hayes

On the last day of February in 1953, according to James Watson, Francis Crick announced to the patrons of the Eagle pub in Cambridge, "We have discovered the secret of life." History supports the boast. If life ever had a secret, the double helix of DNA was surely it. And yet Watson and Crick had *not* laid bare all the secrets of molecular biology. The campaign to understand the code embodied in the double helix was just beginning, and the years ahead would be notable for frustration, false starts and brilliant ideas that turned out to be utterly wrong. It took another full decade to solve the code.

Some weeks ago I found myself browsing in the literature of that curious decade. I had come upon one paper by chance, while looking for something else, and was so intrigued that I tracked down some of the earlier works it cited. A few days later I came back to peel away another layer of references. Then I shifted forward in time to read later summations and histories. (This kind of truffle-hunting in the library stacks is especially engaging when you're supposed to be doing something else.)

What fascinated me about the code-breaking effort was how quickly a biochemical puzzle—the relation between DNA structure and protein structure—was reduced to an abstract problem in symbol manipulation. Within a few months, all the messy molecular complexities were swept away, and the goal was understood to be a mathematical mapping between messages in two different alphabets. The methods for devising codes came from combinatorics; the proposed solutions were judged largely by the criteria of information theory. Efficient storage and transmission of information seemed all-important. The coding theorists were trying to learn the language of the genes, but they might as well have been designing a communications protocol for a computer network.

I was fascinated for another reason as well: Some of the proposed codes were truly ingenious. Indeed, it was hard not to feel a twinge of regret on coming to the end of the story and learning the right answer. Compared with the elegant inventions of the theorists, nature's code seemed a bit of a kludge.

Brian Hayes is a former editor of American Scientist. *Address: 211 Dacian Avenue, Durham, NC 27701. Internet: bhayes@amsci.org.*

What We Didn't Know Then

To enter the world of molecular biology circa 1953, you must first forget all you know. This isn't easy when you come from a world where the sequencing of entire genomes is almost routine, and you can buy custom-made DNA by mail order for 69 cents a base pair.

In 1953 no one had yet read the sequence of bases in any DNA molecule—not one scrap of one gene. For proteins the situation was only a little better. Frederick Sanger was finishing his work on the amino acid sequence of insulin, and a few other fragmentary protein sequences had been published. But the very idea that every protein has a precisely defined sequence, the same in all copies of the molecule, was not yet universally accepted. Even the set of amino acids from which proteins are assembled was still subject to dispute (although Watson and Crick would soon sit down at the Eagle to write out the canonical list of 20). And all the biochemical apparatus for translating DNA into protein awaited discovery. Messenger RNA and transfer RNA were unknown. Ribosomes had been glimpsed in electron micrographs, but their function was unclear.

One area that was not quite so murky was the replication of DNA. From the moment Watson and Crick saw that the four nucleotide bases fit together in specific pairs—adenine with thymine, guanine with cytosine—the mechanism of replication seemed obvious: Unzip the double helix and form two new strands complementary to the original ones. One reason this process was so much easier to fathom was that the replication machinery does not have to consider the meaning of a base sequence in order to duplicate it, any more than a Xerox machine has to understand the documents it copies.

Translation, in contrast, cannot avoid semantics—and yet no one had a clue about how to interpret a sequence of bases. Even the most fundamental questions remained open. For example, since DNA is a *double* helix, should you look for information on both strands? If only one strand

carries the message, how do you know which one it is? And which direction do you read in? Trying to make sense of the genome was like being given a book in a language so unfamiliar you couldn't be sure you were holding it right side up.

The Diamond Code

The first coding scheme inspired by the Watson-Crick structure came from an unexpected quarter. The author was not a biologist or a chemist but a physicist: George Gamow, the chief proponent of the Big Bang theory in cosmology.

In Gamow's initial proposal, which he called the diamond code, double-stranded DNA acted directly as a template for assembling amino acids into proteins. As Gamow saw it, the various combinations of bases along one of the grooves in the double helix could form distinctively shaped cavities into which the side chains of amino acids might fit. Each cavity would attract a specific amino acid; when all the amino acids were lined up in the correct order along the groove, an enzyme would come along to polymerize them.

Each of Gamow's cavities was bounded by the bases at the four corners of a diamond. If the DNA helix is oriented vertically, the bases at the top and bottom corners of a diamond are on the same strand and are separated by a single intervening base; the left and right corners of the diamond are defined by that intervening base and by its complementary partner on the opposite strand *(see Figure 1)*.

Some years later, Crick wrote: "The importance of Gamow's work was that it was really an abstract theory of coding, and was not cluttered up by a lot of unnecessary chemical details...." Actually, Gamow's description of the diamond code had more chemical clutter than many of the later code proposals, but it was indeed the abstract parts of the scheme that made an impression and had a lasting influence. In particular, Gamow's treatment of the problem of mismatched alphabets is still the starting point for textbook accounts of the genetic code.

The alphabet problem is simply that there are 20 kinds of amino acids in proteins but only four kinds of nucleotide bases in DNA. Hence there cannot be any one-to-one mapping from bases to amino acids. Using two bases to represent each amino acid still comes up short, since there are only 16 doublets of bases. It therefore seems that the basic unit of information in the genetic code can be no smaller than a triplet of bases. But there are 64 triplets—more than three times the number needed. Explaining away this excess became a major preoccupation of coding theorists.

Gamow's diamond code—viewed abstractly, after sweeping away the chemical clutter—turns out to be a triplet code in disguise. Although the diamonds have four corners, the paired bases along the horizontal diagonal are complementary, and so only one of them carries any information; the other is entirely determined by the rules that link A with T and C and G. Thus each code word—or "codon"—consists of three bases lined up along one strand. There are 64 possible codons, but not all of them are distinct. Gamow noted that most amino acid side chains are symmetrical, and he therefore postulated that the diamonds could be flipped end-for-end or flopped side-to-side without changing their meaning. For example, the triplet CAG becomes GAC when it is flipped end-for-end, and both of these codons must specify the same amino acid. Flopping CAG side-to-side changes the middle A into a complementary T, so that CTG and GTC are also members of the same family of equivalent codons. When all such symmetries are taken into account, how many distinct codons remain? Gamow counted them up and found the answer is 20—just the magic number he was looking for.

The diamond code had another important property: It was an *overlapping* triplet code. Each nucleotide base (except perhaps at the ends of a strand) claimed simultaneous membership in three adjacent codons. For example, the base sequence GATTACA consists of five overlapping triplets: GAT, ATT, TTA, TAC and ACA. At the time, overlapping triplets seemed like a good idea.

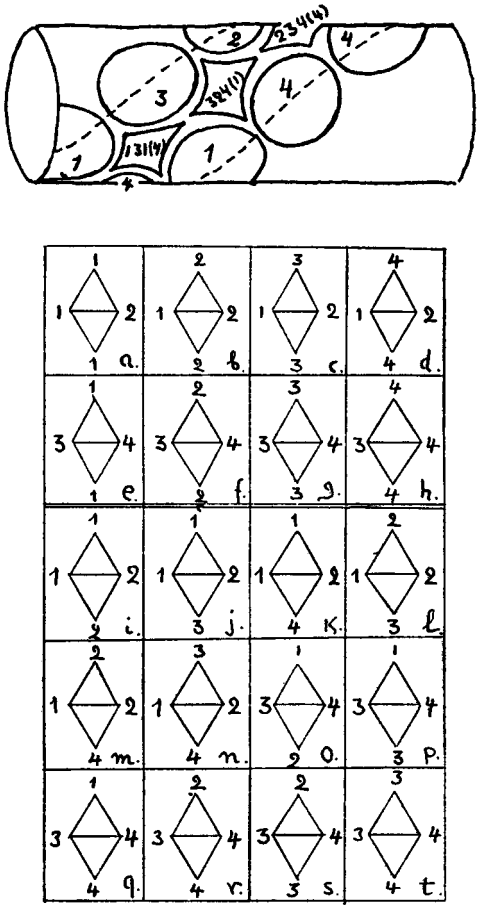

Figure 1. George Gamow's diamond code assumed that proteins form directly on a DNA template. In this 1954 drawing nucleotide bases are designated by numbers and the 20 codons by letters. (Reprinted with permission from *Nature*, 173:318. Copyright Macmillan Magazines Ltd.)

There was a stereochemical justification: The spacing between amino acids in a protein is similar to the spacing between bases in DNA, so that the two polymers mesh best when their subunits are matched one-to-one. The overlapping code also maximizes the density of information storage: Even though three bases are needed to specify any single amino acid, the overall ratio of bases to amino acids approaches 1:1. Finally, overlapping imposes constraints on the possible sequences of amino acids. Gamow thought the constraints might reveal the nature of the code; as it turned out, they were the downfall of his hypothesis.

The RNA Tie Club

A physicist popping up to tell biologists how to solve their problems can't always count on a warm reception. Gamow was welcomed, though, perhaps in part because biology labs in those days were full of carpetbagging physicists. (Crick himself began his career with a physics degree.) Or maybe Gamow just charmed his way in; by all accounts he was an exceptionally amiable fellow. In any case he was soon spending a summer at the Marine Biological Laboratory and collaborating with distinguished molecular biologists. He also founded the RNA Tie Club, limited to 20 regular members (one for each amino acid) and four honorary members (one for each nucleotide base). The ties were wool, with an embroidered green-and-yellow helix. Such an organization might not prosper today—who wears neckties?—but at the time it had an important role in circulating ideas.

The respect accorded to Gamow largely took the form of careful criticism. Attention focused particularly on his overlapping triplets. In any code

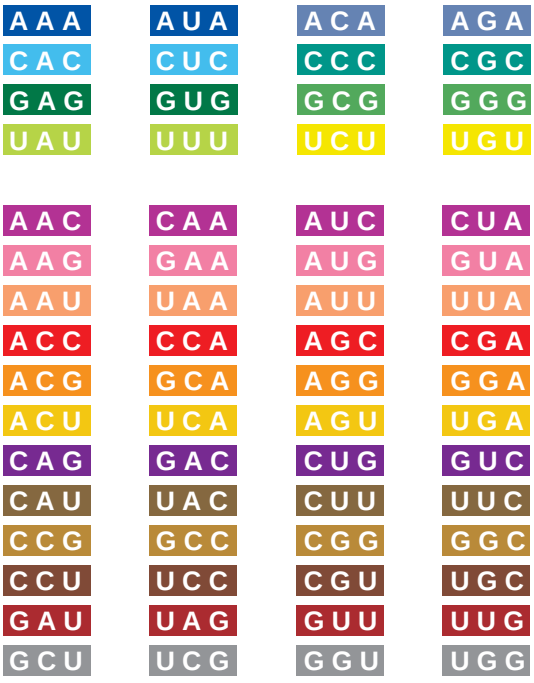

Figure 2. Symmetries of the diamond code sort the 64 codons into 20 classes, indicated here by 20 colors. All the codons in each class specified the same amino acid.

where the ratio of bases to amino acids is 1:1, there are only 4^N nucleotide sequences of length N, but there are 20^N amino acid sequences. It follows that many of the amino acid sequences cannot be encoded by any base sequence. This effect can be seen even in an amino acid sequence of length 2 (called a dipeptide). With 20 kinds of amino acids, there are $20^2 = 400$ possible dipeptides, but two overlapping triplet codons comprise only four bases, so that there are only $4^4 = 256$ combinations. Evidently some 144 dipeptides cannot appear in proteins encoded by an overlapping code.

Even with the sparse protein sequence data available in the mid-1950s, Crick was able to show that the diamond code was ruled out by the experimental evidence. There were known patterns of amino acid repetitions that the diamond code could not produce.

Undaunted, Gamow proposed a "triangle code" that was also overlapping but had different constraints. In this code too the 64 possible triplet codons sorted themselves into 20 families. Later Gamow suggested yet another overlapping-triplet code with an even simpler description: Each codon is defined entirely by its base composition, ignoring the order of the bases within the codon. Thus ACT, ATC, CAT, CTA, TAC and TCA are all members of the same codon family and specify the same amino acid. Remarkably, the number of codon families in this scheme again turns out to be exactly 20. (It is just the number of combinations of four things taken three at a time.)

Still more overlapping codes came from Gamow and his friends. Richard Feynman had a hand in working out one idea. Edward Teller proposed another—a fairly funky scheme in which each amino acid is specified by two bases in the DNA and by the previous amino acid.

But overlapping codes were coming to the end of their string. Patterns of mutations were one source of doubt. With an overlapping code, changing a single base in the DNA could alter three neighboring amino acids, but protein sequence data were starting to show instances of single amino acid replacements. Then came a definitive proof. Sydney Brenner analyzed all the known protein sequence fragments and found enough nearest-neighbor correlations to rule out every possible overlapping code.

In retrospect, the long fixation on overlapping codons seems unfortunate and misguided, but there were strong arguments favoring such schemes. Matching the dimensions of the protein to those of the template seemed important. So did coding efficiency. Natural selection was expected to maximize storage density and avoid any waste of information capacity. Engineers building the computers of the era certainly worked hard to pack in the bits, so why wouldn't nature do the same? No one could have guessed the awful truth—that nature is wildly profligate, that genomes are stuffed with gobs of "junk DNA,"

that storage efficiency just doesn't seem to be an issue except in a few ultracompact viruses.

Still another reason for favoring overlaps was to avoid the frame-shift problem. To understand the nature of this problem, it's best to turn to a very different kind of proposed code—one that I would like to nominate as the prettiest wrong idea in all of 20th-century science.

Comma-Free Codes

By the later 1950s, there was growing support for the idea of messenger RNA—a single-strand molecule acting as an intermediary between DNA and the protein-synthesizing machinery. At the same time Crick was formulating the "adaptor hypothesis," the idea that amino acids do not interact directly with messenger RNA but are carried by small molecules that recognize specific codons. (Today, of course, the adaptor molecules have been identified as transfer RNAs.) The codons were by then thought to be nonoverlapping triplets of bases.

The process of gene expression was imagined as going something like this. First the appropriate segment of DNA was transcribed into messenger RNA; like replication, this was done by blind copying, without regard to the meaning of the sequence. Then the messenger RNA stretched out in the cytoplasm of the cell with its long row of codons exposed like a sow's nipples. Each adaptor molecule, already charged with the correct amino acid, poked around until it latched onto the right codon. When all the codons were occupied, the amino acids were linked together, and the completed protein was peeled off the template.

The scenario must have seemed highly plausible. Even looking back from the 1990s, it seems like the kind of chemistry that living organisms do. The nonsequential pattern-matching needed to line up adaptors on the messenger RNA is vaguely like an enzyme-substrate reaction or like the binding of antibody to antigen. And yet there was a serious problem with the vision of piglets suckling on RNA: A piglet might very well wind up between nipples.

Suppose somewhere in a messenger RNA is the partial sequence … UGUCGUAAG.… (Note that in RNA uracil replaces the thymine of DNA, and so the code is written with U rather than T.) The intended reading is … UGU, CGU, AAG…, but the RNA molecule has no spaces or commas to indicate codon boundaries. The sequence could equally well be read as … UG, UCG, UAA, G … or … U, GUC, GUA, AG.… Each of these alternatives would have a different meaning. Furthermore, in the suckling-pig model of protein synthesis, adaptor molecules that attached to the messenger RNA in different reading frames might interfere with one another and prevent any protein at all from being produced.

The frame-shift problem doesn't arise with an overlapping code, because all three reading frames are simultaneously valid. With sequential

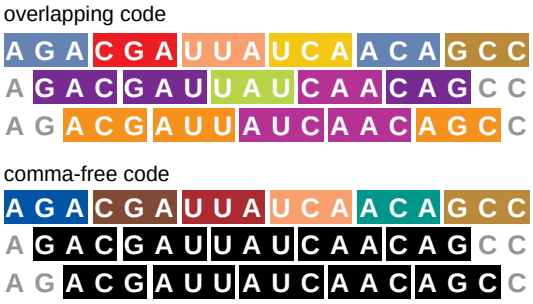

Figure 3. Overlapping code packs 16 codons into 18 base-pairs by exploiting triplets in all three phases, or reading frames. A comma-free code is constructed so that only the codons in one reading frame are meaningful; the overlap triplets are nonsense *(black)*.

codons, however, the translation machinery has to be guided to the right frame. In 1957 Crick devised a solution that seemed at once so clever and so obvious that it just *had* to be right. He suggested that adaptor molecules might exist for only a subset of the 64 codons, with the result that only that subset would be meaningful; the rest of the triplets would be "nonsense codons." Then the trick is to construct a code in such a way that when any two meaningful codons are put next to each other, the frame-shifted overlap codons are always nonsense. For example, if CGU and AAG are sense codons, then GUA and UAA must be nonsense, because they appear inside the concatenated sequence CGUAAG. Similarly, AGC and GCG are ruled out by the sequence AAGCGU. If all the out-of-frame triplets are nonsense, then the message has only one reading. A code with this property is said to be comma-free, since messages remain unambiguous even when words are run togetherwithoutcommasorspaces.

Do such codes exist? In English you might try to find a subset of all three-letter words that can be jammed together without creating any additional instances of the words in the subset. To make the problem more manageable, consider this list of 10 three-letter words: *ass, ate, eat, sat, sea, see, set, tat, tea, tee.* Is there a subset that forms a comma-free language? Trial and error shows that the words *ate, eat* and *tea* cannot all appear together, because *teatea*, for example, contains both *eat* and *ate.* Similarly, *sea* combines with *tat, tea* or *tee* to produce *eat*. One set of words that has no conflicts is *ass, sat, see, set, tat, tea* and *tee.*

How many words can a comma-free code include? For the case of RNA, Crick and his Cambridge colleagues John Griffith (another physicist) and Leslie Orgel carried out a straightforward analysis. They pointed out first that the codons AAA, CCC, GGG and UUU cannot appear in any comma-free code, since they cannot combine with themselves without generating reading-frame ambiguity. The remaining 60 codons can be sorted into groups of three, where the codons within each group are related by a cyclic permutation. For example, the codons AGU, GUA and UAG form one such group. A comma-free code can

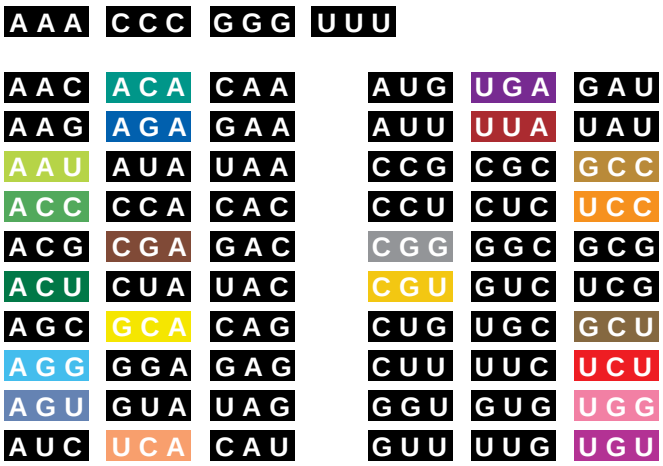

Figure 4. To build a comma-free code, first exclude the triplets AAA, CCC, GGG and UUU, then divide the remaining 60 triplets into groups of three, related by a cyclic permutation. A code can include no more than one triplet from each group.

have no more than one codon from each of these permutation classes. How many classes are there? Dividing 60 objects into groups of three produces exactly 20 groups. Bingo!

The analysis just given sets the maximum possible size of a comma-free genetic code, but it does not guarantee that a maximal code actually exists. Nevertheless, Crick, Griffith and Orgel went on to construct several examples. And they offered a vision of how the code might work: "This scheme … allows the intermediates to accumulate at the *correct* positions on the template without ever blocking the process by settling, except momentarily, in the wrong place. It is this feature which gives it an advantage over schemes in which the intermediates are compelled to combine with the template one after the other in the correct order."

Crick and his colleagues were quick to point out that they had no experimental evidence for the comma-free code. As a nonoverlapping code, it put no constraints on amino acid sequences, so there was no point in looking for confirmation there. The code did strongly constrain the base sequences of DNA and RNA, but those sequences were unknown. "The arguments and assumptions which we have had to employ to deduce this code are too precarious for us to feel much confidence in it on purely theoretical grounds," they wrote. "We put it forward because it gives the magic number—20—in a neat manner and from reasonable physical postulates." The magic number was enough to persuade both biologists and the wider public. Carl Woese later wrote: "The comma-free codes received immediate and almost universal acceptance.… They became the focus of the coding field, simply because of their intellectual elegance and the appeal of their numerology.… For a period of five years most of the thinking in this area either derived from the comma-free codes or was judged on the basis of compatibility with them."

The intellectual elegance also attracted the attentions of coding-theory professionals, most notably Solomon W. Golomb, now at the University of Southern California. Golomb and his colleagues (including the physicist-biologist Max Delbrück) wrote several papers on comma-free codes, taking the biological problem as their point of departure but going on to explore more abstract and generalized ideas. They quickly deduced a formula for the maximum size of a comma-free code: For an alphabet of n letters grouped into k-letter words, the formula takes a particularly simple form when k is a prime: $(n^k - n)/k$. For $n = 4$ and $k = 3$ (the case of interest to biologists) they showed that there are 408 maximal comma-free codes and gave a procedure for constructing them. And they devised some more elaborate related codes. For example, a transposable comma-free code is designed so that both strands of the DNA have the comma-free property. Using triplets, the largest transposable code has only 10 codons, but a quadruplet code yields 20. Golomb also invented a genetic code based on sextuplets; it is not only comma-free and transposable but also can correct any two simultaneous errors in translation, and detect a third error. Life would be a lot more reliable if Solomon Golomb were in charge.

Reality Intrudes

The comma-free codes were not quite the last word in the wildcat era of genetic code–building. In 1959 Robert Sinsheimer suggested a scheme where the genetic alphabet had only two letters; A and C were interpreted as the same symbol, and so were G and U. This device was a way of coping with the recent discovery of wide variations in the ratio of (A + U) to (G + C) in various organisms. Of course reducing the code to binary notation meant that triplets could not code for 20 amino acids; the codons would have to be at least quintuplets (providing 32 combinations).

As far as I know, no one ever proposed a three-letter, ternary code. Such a code might distinguish A from U but lump together C and G, producing 27 codons. This plan has a faint echo in the real genetic code, where the third base in a codon is sometimes interpreted merely as A or G versus U or C.

I'm also surprised that no one gave serious thought to schemes where the codons can vary in length. In engineering, the idea of choosing shorter sequences to represent more frequent symbols was already a well-established trick for compressing a message. David Huffman had created a theory of such codes in 1951, and of course the Morse code went back a century further. Biologists were clearly aware of the principle, and they were mindful of coding efficiency, but they did not explore the possibility.

Perhaps if the era of speculation had continued a few years more, these wrong ideas would also have been given their turn. But in 1961 the whole coding craze was brought up short by unexpected news from the lab bench. Marshall W. Nirenberg and J. Heinrich Matthaei of the National Institutes

of Health announced that artificial RNAs could stimulate protein synthesis in a cell-free system. What's more, the first RNA they tried was poly-U, a long chain of repeating uracil units. In comma-free codes, UUU has to be a nonsense codon, but Nirenberg and Matthaei's result implied that it codes for the amino acid phenylalanine. A few more codons were identified over the next year or two. Then Philip Leder and Nirenberg found an even better experimental protocol, and by 1965 the genetic code was mostly solved.

The code resembled none of the theoretical notions. As the table assigning codons to amino acids was filled in, it became apparent that the magic number 20 held no magic after all. All the clever mathematical contrivances for getting 20 amino acids out of 64 codons turned out to be figments of the human urge to find pattern, not reflections of any natural order. The "extra" codons are merely redundant: Some amino acids have one or two codons, some have four, some have six. (Three codons serve as stop signs.) At first glance the mapping between codons and amino acids appeared arbitrary, even haphazard.

Nature also ignored all the mathematical ingenuity applied to solving the frame-shift problem. The living cell does it by a kind of dead-reckoning. Ribosomes march along the messenger RNA in strides of three bases, translating as they go. Except for signals that mark where the ribosome is supposed to start, there is nothing in the code itself to enforce the correct reading frame.

When I mentioned to a biologist friend that I find some of the hypothetical genetic codes of the 1950s more appealing than the real thing, she protested that the actual code is one of the most elegant creations of biochemistry, and she pointed out some of its subtle refinements. The codon table is not entirely arbitrary. Its redundancies confer a kind of error tolerance, in that many mutations convert between synonymous codons. When a mutation does alter an amino acid, the

Figure 5. Codon assignments show subtle symmetries in Gamow's diamond code, in his similar code that groups together all triplets with the same composition, and in a comma-free code. The actual genetic code appears less regular.

substitute is likely to have properties similar to those of the original. Computer simulations by David Haig and Laurence D. Hurst show that the present code is nearly optimal in this respect.

These observations suggest that I should be grateful my genes were not designed by George Gamow or Francis Crick. With Gamow's overlapping codes, any mutation could alter three adjacent amino acids at once, probably disabling the protein. Comma-free codes are even more brittle in this respect, since a mutated codon is likely to become nonsense and terminate translation.

But criticisms of this kind are not entirely fair. They pluck the invented code out of its theoretical context and plug it into a biochemical system that has been evolving for three billion years or more in concert with a very different code. It's like replacing a man's arms with the wings of a bird and expecting him to fly. The reciprocal transplant would be no more successful. That is, if we should ever visit a planet where life has evolved for a few billion years with a comma-free genetic code, we would doubtless find that our own code was maladaptive.

Imagine that in 1957 a clairvoyant biologist offered as a hypothesis the exact genetic code and mechanism of protein synthesis understood today. How would the proposal have been received? My guess is that *Nature* would have rejected the paper. "This notion of the ribosome ratcheting along the messenger RNA three bases at a time—it sounds like a computer reading a data tape. Biological systems don't work that way. In biochemistry we have templates, where all the reactants come together simultaneously, not assembly lines where machines are built step by step."

The 64-Codon Question

I want to conclude with a question. At the origin of life, the primitive genetic code was surely smaller and simpler than the modern one. It probably included only a few amino acids, or perhaps a few classes of similar amino acids. At some point in its history the code may have functioned as a pure doublet code, ignoring the third base in each codon and specifying no more than 16 amino acids. Then the translation mechanism grew more discriminating, and a few more amino acids were added to the repertory. My question is: Why did this process of differentiation stop at 20 amino acids? There are plenty of spare codons left, and there are other amino acids that need to be gotten into proteins. So why not expand the code further?

One possible answer is that the code is such a vital engine of life that it has been immutable since the earliest stages of evolution. Another answer is that the code *is* evolving steadily toward greater complexity, and we just happened to have discovered it at the 20–amino acid stage. Maybe our descendants will have 60 kinds of amino acids in their proteins. It's worth noting that 20 does not seem to be a hard-and-fast limit. The codon UGA, which is usually a stop signal, sometimes codes for a 21st amino acid, selenocysteine.

A third possibility is that there really is something special about the numbers 64 and 20. The relation can't be the kind of numerological magic invoked by the comma-free codes, but perhaps there is some property of genetic codes that is optimized when the ratio of amino acids to codons approaches 1:3.

Bibliography

Böck, A., K. Forchhammer, J. Heider, W. Leinfelder, G. Sawers, B. Veprek and F. Zinoni. 1991. Selenocysteine: the 21st amino acid. *Molecular Microbiology* 5:515–520.

Brenner, S. 1957. On the impossibility of all overlapping triplet codes in information transfer from nucleic acid to proteins. *Proceedings of the National Academy of Sciences of the U.S.A.* 43:687–694.

Crick, F. H. C., J. S. Griffith and L. E. Orgel. 1957. Codes without commas. *Proceedings of the National Academy of Sciences of the U.S.A.* 43:416–421.

Crick, F. H. C. 1966. The genetic code—yesterday, today and tomorrow. In *The Genetic Code, Proceedings of the XXXI Cold Spring Harbor Symposium on Quantitative Biology*. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory of Quantitative Biology, pp. 3–9.

Crick, Francis. 1988. *What Mad Pursuit: A Personal View of Scientific Discovery*. New York: Basic Books.

Gamow, G. 1954a. Possible relation between deoxyribonucleic acid and protein structures. *Nature* 173:318.

Gamow, G. 1954b. Possible mathematical relation between deoxyribonucleic acid and proteins. *Det Kongelige Danske Videnskabernes Selskab, Biologiske Meddelelser* 22:1–13.

Gamow, George, Alexander Rich and Martynas Yčas. 1956. The problem of information transfer from nucleic acids to proteins. *Advances in Biological and Medical Physics*, Vol. 4., pp. 23–68. New York: Academic Press.

Golomb, S. W., Basil Gordon and L. R. Welch. 1958. Comma-free codes. *Canadian Journal of Mathematics* 10:202–209.

Golomb, S. W., L. R. Welch and M. Delbrück. 1958. Construction and properties of comma-free codes. *Det Kongelige Danske Videnskabernes Selskab, Biologiske Meddelelser* 23 (9):1–34

Golomb, S. W. 1962. Efficient coding for the desoxyribonucleic channel. *Proceedings of Symposia in Applied Mathematics*, Vol. 14, Mathematical Problems in the Biological Sciences, pp. 87–100. Providence: American Mathematical Society.

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

Judson, Horace Freeland. 1996. *The Eighth Day of Creation: Makers of the Revolution in Biology*. Expanded edition. Plainview, N.Y.: Cold Spring Harbor Laboratory Press.

Leder, Philip, and Marshall Nirenberg. 1964. RNA codewords and protein synthesis, II. Nucleotide sequence of a valine RNA codeword. *Proceedings of the National Academy of Sciences of the U.S.A.* 52:420–427.

Nirenberg, Marshall W., and J. Heinrich Matthaei. 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences of the U.S.A.* 47:1588–1602.

Sinsheimer, Robert L. 1959. Is the nucleic acid message in a two-symbol code? *Journal of Molecular Biology* 1:218–220.

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