The genetic code by numbers

The application of the theory of mathematical groups to the origin of the genetic code will startle molecular biologists, but is best regarded as a valuable exercise in classification.

THE problem of the genetic code has several facets, of which the most compelling is that of why it is just what it is. Once Crick and Brenner had established that the code is a triplet code, with each consecutive triplet in a molecule of either DNA or RNA corresponding to a single amino acid in the ultimate protein (*Nature* **192**, 1227; 1961), it was natural that people should look for an explanation, both for its own sake and because an understanding of how the code evolved must certainly be a pointer to the origin of life itself. It was already clear that the genetic code is not merely an abstraction, but also the embodiment of life's mechanisms; the consecutive triplets of nucleotides in DNA (called *codons*) are inherited but they also guide the construction of proteins.

So it is disappointing, but not surprising, that the origin of the genetic code is still as obscure as the origin of life itself. That is not to say that people have not been worrying about the problem, although most of the worrying seems to have been done at an early stage, in the 1960s and early 1970s, soon after the decryption had first been done. Is that a mark of people's disappointment, or merely a sign that they have had better things to do?

Historically, there have been two views of the origin of the code. The most obvious, that there must be some physical chemical interaction between a nucleotide triplet of the code in a nucleic acid polymer represented by a molecule of messenger-RNA (mRNA) and the amino acid it specifies in the eventual protein, was first advocated by Carl R. Woese and his associates. That was knocked down by the eventual demonstration that amino acid molecules are scavenged from the cytoplasm of a cell, and then carried to the ribosomes at which protein molecules are assembled from amino acids, by RNA molecules called transfer-RNA, tRNA for short. Although tRNA molecules are small as RNA molecules go, there is no obvious way in which the amino acid molecules they carry can interact with their signatures, the *anticodons*.

A little definition will help. The DNA code has four elements, called T, C, A and G (for thymine, cytosine, adenine and guanine). The RNA is similarly specified except that U (for uracil) replaces T. In double-stranded DNA, A and T are paired together in oppositely directed strands, as are G and C. In RNA molecules, tRNA molecules in particular, physical pairing of this kind is as often internal, helping to shape the secondary structure of the molecule. The characteristic anticodon of a tRNA molecule is the ribonucleotide triplet that is complementary to the triplet of the mRNA specifying a particular amino acid. For example, if UGC is an RNA codon (which happens to specify the amino acid cystine), the anticodon will be ACG, invariably found unencum-bered by internal pairing. The cystine molecule is then attached to the other end of the tRNA.

Long before that was known, some general features of the genetic code were plain. For one thing, there are 4 × 4 × 4 = 64 possible codons, but only 20 amino acids are known naturally to occur. Even allowing for the need to specify the points at which the transcription of DNA into RNA normally stops, the genetic code must be redundant; some amino acids must be specified by many different codons. As early as 1961, Crick and Brenner put their money on degeneracy (rather than on the alternative that the superfluous codons meant nonsense). Among other things, they saw it as a way of explaining how microorganisms whose nucleotide composition is very different could have proteins with similar amino acid composition.

By 1968, Crick had carried the argument a step further (*J. molec. Biol.* **33**, 367; 1968) by demonstrating with cartesian logic that if the genetic code is now a triplet code, it must always have been a triplet code: a change in the base would always have meant the mistranslation of essential proteins and thus a loss, perhaps catastrophic, of darwinian fitness. Crick also put his weight behind the notion that the primitive genetic code would have been less specific than that which now holds sway, either specifying fewer amino acids or specifying classes of amino acids, perhaps defined by their acidity.

Much less has happened since 1968 than might reasonably have been expected. Perhaps the landmark is the paper by T. H. Jukes in 1983 (*J. molec. Evol.* **19**, 219; 1983) in which he used a comparison of the universal genetic code (applicable in all cells) and that in mitochondria (then just recognised to be different) to argue that the primaeval genetic code must have used sixteen anticodons to code for fifteen amino acids (with the extra anticodon used as a stop signal). That does at least point out that, when more is known of the dynamics of the genome, it should be possible to unravel the evolution of the code from the regulation and placing of tRNA genes and from the properties of defective pseudo-genes.

Meanwhile, quite a different way of regarding the evolution of the genetic code has come to light, one likely to be more familiar to particle physicists than to molecular biologists. In the last issue of *Physical Review Letters* (**71**, 4401; 1983) for 1993, José Eduardo M. and Yvone M. M. Hornos, both from the University of São Paulo, argue that the problem of the genetic code is simply a problem in symmetry breaking similar to the reasons for the supposed existence of, say, the Higgs boson, and therefore best described by group theory.

The argument is not negligible. The starting point is the recognition that there are 64 different codons (or anticodons). If all the amino acids were once the same (or if the differences between them did not matter), that would be a highly symmetrical circumstance. But if differences matter, then the symmetry is destroyed or broken. And then it is permissible to look for chains of abstract mathematical groups which, if multiplied together so that the elements of the product group combine in ways determined by combinations of each component group, will reflect the classifications observed in the real genetic code.

The simplest analogy is with, say, the allowable electron states in the second principal quantum level of an atom — the line of the periodic table beginning with lithium. There are four of them, meaning that the electron shell concerned can accommodate eight electrons when allowance is made for the two-valued spin of the electron. But ordinarily, the existence of four states is not apparent. Only when there is a magnetic field does it become plain that one of these states has zero angular momentum, that the other three have unit angular momentum and that the latter three may be oriented in three different ways with respect to the external magnetic field. The textbooks are full of these, the nearest things to cladograms in atomic physics. They show how energy levels split into sub-levels under the influence of asymmetry.

So, it seems, it may be with the genetic code. The Hornoses have found a mathematical group with a 16-dimensional representation (the symmetric primaeval genetic code) which can be broken down into a product of simpler groups reflecting the pattern of redundancies observed. They make the system work, and even conclude that there may have been an intermediate code specifying only 15 amino acids, just as Jukes suggested. But sadly, even if the argument were much more compelling, only laboratory work can settle what really happened. That remains a challenge.

John Maddox