Loose ends

by Sydney Brenner

Molecular biology by numbers ..... three

Three nucleotides correspond to each amino acid in the genetic code and triplets were thought to be the coding ratio from the very early days. The key to all the work was to define the *proper* amino acids — that is, those found in proteins — and exclude all the others. In the 1950s, textbooks of biochemistry, vying with each other for the length of the list of amino acids they could produce, included both cysteine and cystine, citrulline, ornithine and even *D*-amino acids, β-alanine and hydroxy-proline. Getting the twenty was the fundamental step. Francis Crick and Jim Watson wrote them down and I got to that number from the fragments of peptide sequences being accumulated by Fred Sanger and others. Three nucleotides was the minimum that could be used to get twenty, although there was a suggestion, which Francis called the 'naive biochemist's code', that sixteen of the twenty amino acids were coded by doublets and the remaining four by singlets. (This has an echo in a phrase I often use — 'the naive molecular biologist's gene', which is almost exactly one kilobase long because NMBs believe that all proteins are exactly 333 amino acids long.)

What made the early days of genetic cryptography difficult was the self-imposed stereochemical constraint. The 3.3 Å repeat of nucleic acids is about the same as the 3.5 Å chemical repeat of the polypeptide chain, and it was thought that this one-to-one physical correspondence was necessary for the mechanism of protein synthesis. Maintaining a coding ratio of three nucleotides and a step size of one nucleotide for each amino acid requires special solutions which were first clearly stated in Gamow's 'diamond' code. This was the first of the overlapping triplet codes, in which, in a nucleic acid string, nucleotides 1,2 and 3 code for the first amino acid, nucleotides 2,3 and 4 code for the second and so on. The diamond code was degenerate in the sense that more than one triplet corresponded to a particular amino acid; some had four, and others, two. The particular decomposition was obtained by the application of a rule based on totally implausible and unrealistic physical assumptions.

Gamow's particular code could be disproved for the known protein sequences, but it had already become clear that there were many ways of degenerating the triplets and, being biology, it could have been an accident of evolution and quite arbitrary, rather than being derived from some elegant mathematical rule. It was not feasible to test all of the codes, one by one, for compatibility with the data. Indeed, there was a paper to show that if we were going to do this by computer we needed one several orders of magnitude more powerful than those available at the time and that we should have started the work at the onset of the fall of the Roman Empire.

I realized that all overlapping triplet codes had one thing in common, regardless of the degeneracy. Because a dipeptide would be coded by four bases, the codes all constrained the number of possible dipeptides to 256, rather than the 400 that are the maximum number of dipeptides possible from 20 amino acids. There were insufficient data to test this prediction directly but by the autumn of 1954 I had statistical evidence that the known dipeptide occurrences fitted a Poisson distribution based on 400 rather than 256. I showed my chart to Gamow who promptly lifted it and put it in a review he was writing with a footnote acknowledging that I had done it as well. Coding was therefore not only my first sally into theoretical biology but also my first encounter with conduct in modern science. Shortly thereafter, I found the proof that all overlapping triplet codes were impossible and this led to the sterochemical constraint of one nucleotide – one amino acid being discarded. I proposed what I privately called the 'Humpty-Dumpty' theory of protein synthesis, which was that it begins at the beginning, goes on until it reaches the end and then stops. When Francis proposed the 'adapter hypothesis' we knew that the code would only be found empirically, not through the exercise of the mind. Theoretical coding died and its most interesting product, the elegant comma-less code, became an historical curiosity.

The other triad in molecular biology is embodied in the 'central dogma', often expressed in Middle Sloganic as DNA makes RNA, RNA makes protein. I have always been slightly puzzled why Francis chose the word 'dogma' as he is the last person to be described as a church man, even of the most liberal and reformed kind. When reverse transcription was discovered, many people gleefully tried to depose the central dogma but, as has been made clear by Francis, the rule really applies to nucleic acid and proteins; that there are two kinds of nucleic acids and ways of going backwards and forwards between them is trivial. The dogma is better and more deeply stated in the diadic form: once information gets out of DNA into protein it cannot go back again. Perhaps it became a triad because if there is a beginning and an end there has to be a middle.

I once formulated the 'central dogma of biotechnology' as DNA makes RNA, RNA makes protein, and protein makes money. For this, I won an exceedingly small prize in Japan and my work was translated into one language. Only later did I realize that I had missed a golden opportunity to increase my compensation. Introducing the fourth component breaks the original dogma and allows closure of the cycle, because money allows information to be taken out of protein and put back into DNA. That is what we are all doing nowadays and, if one wants an anthropic principle in science, this is a much better one than that talked about in cosmology. Money does make the world go round.