Loose ends

Molecular biology by numbers ..... one

by Sydney Brenner

Classical geneticists considered the gene to be one indivisible unit of mutation, recombination and function. The picture was that of beads on string, with recombination taking place between the beads, and mutations creating altered bead states or alleles with different functions. How genes worked was a mystery, but the one gene–one enzyme hypothesis formulated by Beadle made the correct connection — although as late as 1954 there were still people who thought that genes could make carbohydrates or even phospholipids. The unitary hypothesis began to show cracks even before the discovery of the structure of DNA, when rare recombination events had been found between mutations in what was thought to be one gene. These had to be called pseudoalleles and there were even dark hints of the existence of subgenes.

It was Seymour Benzer's work on the fine-structure genetics of the *rII* locus in bacteriophage T4 that destroyed the classical model of the gene. He showed that each gene, defined as a functional unit, contained hundreds of mutational sites that could be separated by recombination. A simple calculation revealed that this map was on a scale that corresponded to individual base-pairs of DNA.

To mark this new view of the gene, Seymour invented new terms for the now different units of mutation, recombination and function. As he was a physicist, he modelled his terms on those of physics and just as electrons, protons and neutrons replaced the once indivisible atom, so genes came to be composed of mutons, recons and cistrons. The the unit of function, the cistron, was based on the *cis–trans* complementation test, of which only the *trans* part is usually done.

Of these terms, only cistron came to be widely used. It is conjectured that the other two, the muton and the recon, disappeared because Seymour failed to follow the first rule for inventing new words, which is to check what they may mean in other languages. In his case it was French that did him in; muton is far too close to the word for sheep, and recon can be confused with an insult used by taxi drivers in Paris. Incidentally, I was told of another example of this principle in the form of an anti-freeze spray used on car doors in the winter in Finland that had a name very much like.Piss.

Seymour's pioneering invention of units was followed by a spate of other new names, not all of which will survive. One that seems to have taken root is codon, which I invented in 1957; and the terms intron and exon, coined by Walter Gilbert, are certain to survive as well. Operon is moot; it is still frequently used in prokaryotic genetics, but as the weight of research shifts to eukaryotes, which do not have such units of regulation, it may be lost. Replicon, invented by Francis Jacob and myself in 1962, seems also to have survived, despite the fact that we paid insufficient attention to how it sounded in other languages. This struck me forcibly some years later when a Japanese colleague asked me what I thought about the leprechaun hypothesis.

Units are needed in science whenever measurements are made. Physics has dozens of them named after physicists. There are Ångstroms, newtons, joules, einsteins, debyes, curies, and so on. We have svedbergs in biology, but sedimentation constants are still close to physics. There is, of course, the centimorgan for the measure of recombination, but I think we could do with more. Perhaps we should get rid of kilobases and substitute kilowatsons, and substitute crick for triplet. We could then say that the human genome has 3 000 megawatsons (or 3 gigawatsons) of DNA, and the average coding sequence in eukaryotes is 410 cricks long. And, of course, for those of us who study evolution, one million years must be called a darwin.

I have used the word quit as the logarithmic unit of sequence information. Thus, a bacterium with 4 megawatsons of DNA could be said be contain 11 quits of sequence information ($4 \times 10^{6} = 4^{11}$). The careful reader will notice that one quit equals two bits, and that the human genome, with 16 quits of sequence information, makes the human a 32-bit animal.

I have been struck by the fact there is no unit for the unit. I am an assiduous collector of errata, and I recently found a gem tucked away in a corner of *Nature*, urging the reader to substitute the words "500 micrograms" for "500" and "25 millilitres" for "25" in what must have been a mysterious paper. At the time, I also realized that this provided a wonderful way to delay the work of one's scientific competitors. Just imagine the erratum that says for "kilograms" read "micrograms". I had thought that these and related problems could be solved if we had a special word for the unit itself. I toyed with the idea of using cantor or piano, or even frege, from the realms of the theory of arithmetic, and it took me some time to realise that we had a better one closer to hand. I therefore propose that we use the word monod as the unit for the unit. Instead of saying 125 millilitres we would say 25 millimonods of litres; and instead of 128 nanoseconds we would say 128 nanomonods of seconds.

Sadly I have just discovered that the word monod is easily confused with the word for idiot in a Sudanese dialect, so this will not work.