Long-range correlations within DNA

The idea that nucleotide bases in strands of DNA may be correlated over several thousands of base-positions has been established, but has not yet been explained.

ARE there long-range correlations between the nucleotides within a gene and, if so, what purpose can they serve and how, in any case, do they arise? These are the questions undoubtedly prompted in many minds by the publication in March this year, by C.-K. Peng of Boston University and his colleagues (*Nature* **356**, 168; 1992), of an article demonstrating long-range correlations extending over many thousands of base positions in several different DNA sequences. It was natural to expect that the noise level would soon be filled with the clash of opinions between sceptics and believers, no doubt equipped with at least one supposed cause of error, or with an explanation as the case might be. Instead, surprisingly, it is the silence that has been deafening.

So far, at least. But now Richard F. Voss from the IBM Research Center at Yorktown Heights, New York, has responded in *Physical Review Letters* with what may be a more explicit way of calculating the correlations (**68**, 3805; 22 June 1992). The results are inherently the same — there are indeed long-range correlations along the length of real-life stretches of DNA. The long-range correlation may even differ significantly between taxonomic groups. But nothing is yet said about their origin or function.

The expectation that there will be no long-range correlations is natural, and is the obvious baseline against which correlations in real DNA are measured. That there may be short-range correlations in DNA is another matter; crude expectation would suggest there might well be. Armchair molecular biologists will quickly suggest several plausible patterns: purines (adenine and guanine) may alternate with pyrimidines (thymine and cytosine), purines (or pyrimidines) may alternatively predispose to successors along the chain of their own kind, or these patterns may occur in different parts of a gene according to its function. But nobody would expect that correlations such as these hypotheses imply would extend over more than a few dozen bases, corresponding to nearest neighbour interaction between nucleotides along the length of a twisted double helix.

So how to calculate the chance that the illusion of long-range correlation will arise by pure chance? By the calculus of the random walk, as used in Einstein's theory of Brownian motion. If, for example, the issue is simply whether there is a correlation between the occurrence of a purine base at one point in a stretch of DNA and a purine (or a pyrimidine) base 10,000 base-positions further along, the first step is to know what chance would throw up. One could imagine (as Peng *et al.* did) the structure of a particular gene plotted on a one-dimensional graph — one step to the right for a purine, one step to the left for a pyrimidine. Then the problem is simply that of Brownian motion in one dimension.

If chance is all that matters, the result is straightforward. The likelihood that there will be a purine here and also a purine there will decrease steadily (monotonically, as they say) with the distance between here and there. For a strictly random walk, the square root of the average of the square of the displacement of the point on the walking graph from its average position (the root mean square displacement) will be proportional to the distance to the power 0.5. Another way of putting that (after a little algebra) is that there will be a gaussian distribution of the position of a particle along a one-dimensional lattice which becomes more widely spread out as time passes (or as the distance between base-positions increases). Any other kind of behaviour implies that some nonrandom process is under way.

So how best to test that the long-range correlations between nucleotides in DNA differ meaningfully from these expectations? There are obvious pitfalls. To deal separately with the four bases that crop up in real DNA, it is essential to consider the process of random-walking in four dimensions; otherwise, correlations that are artefacts will crop up. And what happens when there are external constraints in the analysis of a particular stretch of DNA, as when it is richer in C and G than in A and T (or conversely)?

Voss's way has an intrinsic simplicity, guided by the binary character of the arithmetic of the computer. Represent the sequence of a gene by four separate sequences, each referring to one of the nucleotides that ordinarily crop up; the sequence contains a '1' if a site is occupied by that nucleotide concerned, otherwise a zero. Adding the four sequences together then gives a simple sequence of '1's. From these, it is then possible to calculate all possible correlations for a given DNA sequence, say the chance that there will be an 'A' at one position and a 'C' exactly fifty positions further along the strand, and to take the average over all possible placings of the first position. Readers of Voss's beguiling article should be warned that there is some fancy binary algebra embedded in his proof of the link between the elementary sequences and that from the real world.

Voss's biggest single achievement is to have run the complete genome of cytomegalovirus (strain AD169) through his computer, but altogether he has analysed more than 50 million nucleotide position in the more than 25,000 DNA sequences distributed as GenBank Release 68. The outcome is quite remarkable. There are indeed long-range correlations, extending over thousands or more of base positions.

Voss hammers home the point that the character of the long-range correlations is that of what is called '1/*f* noise', which is shorthand for the notion that a noise-level may be a fractional power (not very different from unity) of 1/*f*, of the kind that arises in systems of a fractal character among others. But using the fractional power as an index (say β), it emerges that there are significant differences (ranging from 0.76 to 0.92) in the values of β describing the long-range correlation between the occurrence of the same bases.

In passing, the speculations of the armchair molecular biologists are confirmed. In cytomegalovirus, pyrimidines are anti- correlated at some sites while at others the opposite is the case. One intriguing feature of the analysis is that the triplet DNA code shows up as a bump in what would otherwise be the 'high- frequency' end of the spectrum where the frequency corresponds to a spacing of three units. It is also interesting that Voss has used as a random 'control' a sequence consisting of the first 1.13 billion decimal digits of the number π.

The evolutionary comparisons of long-range correlations are the most intriguing. Both phage and bacteria have β roughly equal to 1.0, but the spectral density does not vary linearly (on a long-log plot) as for other taxonomic groups. Vertebrate sequences yield a similar slope, but a more regular one. The lowest values of β are for primates and organelles (such as mitochondria). In all groups,there is a peak of short-range correlation corresponding to a three-nucleotide displacement, but in vertebrates, invertebrates and primates (but not rodents or mammals in general) there is also a peak corresponding to a displacement of nine units.

Voss is modest about his work, describing it as "only a beginning". He can say that again. No doubt others are already preparing to follow where he has begun. But still, there is no explanation of these long-range correlations. Are evolutionary processes at work? Or may the long-range correlations arise because genes in higher organisms have been derived by patching together genes from simpler entities? **John Maddox**