MATHEMATICAL BIOLOGY

DNA Shows Unexplained Patterns Writ Large

A tourist experiencing Mount Rushmore solely by pressing his nose up against the mountain wouldn't notice that many of the patches of rock right in front of him were actually part of a huge sculpture of presidential faces. Put him a mile away, though, and he'd instantly see the collective structural meaning of the various patches. Many molecular biologists have been rather like those short-sighted tourists, peering at immense DNA molecules from vantages so close that they can only clearly see stretches of tens or hundreds of nucleotides at a time. But a few mathematically inclined investigators have recently stepped back to take the long view of the molecules' nucleotide sequences. And some large-scale molecular patterns have leaped out at them.

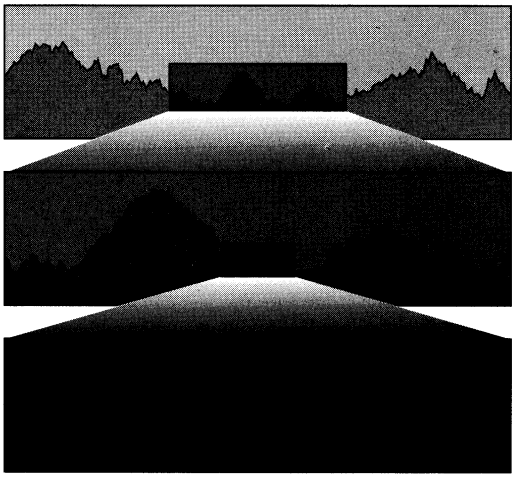
The patterns are far more abstract than George Washington's profile on a mountainside, but just as surprising: long-range correlations in the DNA sequence. A specific nucleotide at one site in a DNA strand appears to have a bearing on which of the four possible nucleotides (A, C, T, and G, for short) occupies a specific site 100, 1000, 10,000 or even more nucleotides away. For example, a T at one site might slightly increase the chance that an A will appear 8214 sites away. To put it in meteorological terms, it's as if sunny weather at 2 PM every 14 January tended to be correlated with fog at 9 AM every 8 September.

And it's just as unexpected. Molecular biologists are used to shorter-range patterns in nucleotides, which come in standard sets of three, each set coding for one amino acid of a protein or serving as a regulatory signal in gene transcription. But "it is almost incredible that the occupant of one site on a gene would somehow influence which nucleotide shows up even 100,000 bases away," remarks H. Eugene Stanley of Boston University, who collaborated with C.K. Peng, Sergey Buldyrev, Shlomo Havlin, and Francesco Sciortino of Boston University and Ary Goldberger and Michael Simons of Harvard Medical School on one of the first papers to announce the correlations.

That leaves researchers at sea about the significance of these correlations. "It's meaningful, but just what it means isn't at all obvious," says Charles DeLisi, the dean of the College of Engineering at Boston University, who is credited with being the first to propose a human genome project when he was at the Department of Energy.

Hints of the large-scale patterns were first reported last February by Wentian Li of Rockefeller University and K. Kaneko of the University of Tokyo. To the gene for a coagulation factor in human blood, they had applied a "mutual information function"—a measure of the influence that one symbol on a string of symbols, say a letter in a string of letters or a nucleotide in DNA sequence, has on the identity of another symbol elsewhere in the string. Li had already been running other DNA sequences through the function for more than 3 years, searching for long-range correlations, but had turned up nothing unusual. But when he and Kaneko tried the coagulation gene, up popped some unexpected correlations between distant nucleotides, as they reported in the 7 February *Europhysics Letters*.

With a sample of just 12,850 bases, though, Li and Kaneko were still peering at DNA on too small a scale to make out the full pattern. But the next month, in a paper in the 12 March *Nature*, the Boston collaboration unveiled an analysis of a variety of genes and DNA sequences, based on another correlation-seeking method called the DNA walk. The researchers checked to see which class of nucleotide—pyrimidine or purine—is present at successive positions. When they found a pyrimidine (T or C), they took a step forward in their imaginary DNA walk; when it was a purine (A or G), a step backward. In a random sequence of purines and pyrimidines, such a walk would tend to leave you where you started after you had examined thousands of nucleotides. If correlations existed, but only over short ranges, you would stroll back and forth across your starting point. But long-range correlations would take you some distance away.



SOURCE: H.E. STANLEY, *ET AL.* ILLUSTRATION: C. FABER SMITH

A fractal DNA landscape. A "DNA walk" reveals similar patterns on different scales.

And that's what happened with many of the 24 viral, bacterial, yeast, and mammalian sequences the Boston group analyzed for their *Nature* paper, as well as the additional 60 or so they've looked at since then. What's more, the Boston septet found that the degree of correlation remained the same over distances of 1000, 10,000 or 100,000 positions. To physicists and statisticians, that scale-invariant behavior is a sign that the complex phenomenon under study has fractal properties—the statistical patterns emerging in DNA sequences, it seems, are like a coastline on which the same pattern appears whatever the scale on which you view it.

But what does all this mean? For IBM physicist Richard Voss, who in the 22 June *Physical Review Letters* reported a still larger analysis of 25,000 genes—amounting to more than 50 million nucleotides—this so-called self-similarity is titillatingly reminiscent of that seen in other fractal phenomena, including music, stock market fluctuations, and traffic flow. But he admits that for biologists, "this may end up being garbage"—nothing more than a curiosity.

That outcome seems unlikely to Stanley and Goldberger. The studies done so far, they say, offer some tantalizing hints about what might underlie these long-range correlations. In the Boston study, for example, the genetic sequences that did not reveal long-range correlations all had something in common: They lacked the mysterious noncoding stretches of DNA known as introns. Introns interrupt most genes in higher organisms, but not those of bacteria and cellular organelles. The apparent importance of introns explains Li's initial failure to find correlations: His first samples of DNA lacked introns, but the blood coagulation factor gene—where the correlation finally turned up—is 76% introns.

Those findings hint that a solution to the mystery might lie buried in the introns. "It always struck me as unlikely that introns are junk," says Goldberger. But any answer may also have take into account another finding, from Voss' study: When he compared the correlations that occur in DNA from the 10 different genetic categories in his sample, including primates, invertebrates, and organelles, he found that they exhibited slightly different fractal behavior. On the largest scales, it seems, DNA from each class of organisms has its own statistical signature.

Not that those clues narrow the range of possible explanations for the correlations by much. Among the possibilities Goldberger cites: unknown evolutionary principles, clues to unrecognized chemical and physical constraints on DNA replication, or something else entirely. As Stanley notes wryly, "This work raises more questions than it answers."

–Ivan Amato