Complicated measures of complexity

The search for a means of telling the complexity of numerical data has been urgent but frustrating. Now, there may be some relief in the assertion that there is no single measure.

EVERYBODY wants to be able to measure complexity, but nobody quite knows how it should be done. That is the impression one is likely to form by skimming through the literature on the subject, which is understandably spilling out in the most unexpected places — mathematics journals, those concerned with computer science and communications and straightforward physics journals as well as, inevitably, in Wolfram's *Complex Systems*.

The issue has evidently been made pointed by the recent wave of interest in deterministic chaos. But why is it so difficult?

The essence of the problem is neatly put by Peter Grassberger of the University of Wuppertal (*Int. J. theor. Phys.* **25**, 907; 1986) in these terms. Imagine three two-dimensional patterns, one of them an array of black and white squares as on a chessboard, one the now-familiar kind of trajectory traced out in two dimensions by a chaotic system — typically an intricate pattern of nearly periodic orbits grouped together into whorling shapes — and the third simply a random pattern of dots. Both the first and the third are inherently simple. A few simple rules will help one to construct a chequerboard, while a random pattern may need even fewer rules. The chaotic pattern, on the other hand, can be reproduced only by simulating the system that generates it. Intuitively, it is by far the most complex. But this is not what the numbers say, or at least some of them.

To the extent that many measures of complexity boil down to specifying the information content of a pattern, they turn out to be the equivalents of entropy. And entropy, of course, is greatest for disordered systems. The entropy of the random pattern is thus greater than that of the chaotic pattern, which in turn is greater than that of the chequerboard. Such measures of complexity, including those derived from Shannon's theory of communications systems, all give pride of place to the random pattern of dots, which offends expectation.

The standard answer may be counterintuitive, but the antecedents of this question are interesting in themselves. To make progress, one needs a way of generating patterns with more or less complexity. To make life simple, it is best to begin with one-dimensional patterns when, 'without loss of generality' as they say, one can conveniently think of a pattern as consisting of an infinite string of symbols which may, to suit the computers, be either '0' or '1'. To begin with, one needs a way of generating such patterns.

Grassberger's article turns out to be a neat and intriguing review of much that has been done to define measures of complexity. It is intriguing that he quotes two essays in biology — an attempt to define the most economical taxonomic tree, and to describe life mathematically — as part of the inspiration of his own account. (The second of these authors, G. J. Chaitin, is cited as the source of an interesting view on the meaning of randomness in a piece of Scientific Correspondence we shall be publishing next week.)

Constructing symbol strings with simple structure is easy. By alternating '0's and '1's, one generates simple strings such as '01010101...'. Generating more complex patterns requires more ingenuity — one might wish to generate patterns in which the sequence '...111...' never appears, for example. That, it turns out, can be done by constructing an appropriate directed graph, which is nothing more than a set of points (nodes) on a plane which are joined to their neighbours by lines which carry arrows to show the direction in which movement is permitted. Each line is also marked by one of the symbols making up the string, either '0' or '1'. The rule is to find the starting point (which may have only two arrows leading away from it), to flip a coin at each node at which one arrives and to take as the next symbol in the growing string the symbol on the line chosen as a consequence.

Take a simple square, for example, with all the arrows following in the same direction, but with '0's and '1's alternating around the square. The sequence generated is simply '010101...' with an appropriate choice of corner at which to start. But if one adds a single diagonal from the same corner, but carrying the symbol '1', sequences such as '0101' and '101' will be mixed in the string at random. Much more complicated patterns can easily be generated. The connection with chaotic systems can be made direct by using a generating function, of the kind that chaos-mongers live and breathe — something such as $x_{i+1} = f(x_i)$. The idea is that this equation is a way of indefinitely remapping the points in some interval of the x axis onto the same straight line.

The practitioners in chaos are usually more interested in mapping some finite interval — say that between 0 and 1 — onto itself, which is convenient for pattern generation because it is then possible to assign to x_{i+1} the value '0' if it is less than 0.5 and the value '1' if it is 0.5 or greater.

One of the most intriguing features that keeps cropping up in this literature of pattern formation and analysis is that there are repeated references to Noam Chomsky, who is largely responsible for putting the grammar of language on a logical basis. The reason is straightforward — strings of two symbols with restrictions, such as '...111... is disallowed' can be linked to strings of letters, or words, and the restrictions to the rules of grammar. Indeed the set of all the algorithms by which the patterns are generated becomes the set of all the rules of an infinite set of grammars. In language, the absorbing question is the degree to which rules like these represent real language.

That may seem a diversion, but it is an important one. Another is with the theory of computation and the Turing machine, which is a way of relating the complexity of a computational last to the characteristics of the ideal computer required to carry it out.

Grassberger, for what it is worth, defines four measures of complexity for strings like these, of which he says he believes the most relevant for physical problems is that called *effective measure complexity*. He goes on to show that the definition allows for the complexity to be infinite which the entropy is zero.

But now, it seems, a more radical view has come to the surface. G. d'Alessandro and A. Politi of the Instituto Nazionale di Ottica at Florence have made a direct attack on the difficulty that random patterns are assigned great numerical complexity by a quite novel procedure (*Phys. Rev. Lett.* **64**, 1609; 1990).

What they imagine is that the complexity of a string will eventually be measured by a hierarchy of numbers, the first of which is akin to an entropy and defined as the logarithm of the number of admissible sequences (given the grammar) of length n divided by n as the numbers become very large. But why not do the same for the forbidden words, deriving a second measure in exactly the same way? Out of that definition tumbles directly the notion that the complexity of a random string is identically zero. Their next task is to generalize the argument to several dimensions. One is naturally left wondering why it should ever have been thought that there could be a single number measuring complexity of very different kinds.

John Maddox