In this issue

The language of DNA: How to understand what a sequence tells you about itself

Reduced to its elementary building blocks, a piece of English text is just a 'character string' – a chain of 50-odd symbols (alphabets, digits, punctuation marks, etc.) arranged in a sequence, some symbols occurring more often than others. A sonnet by Shakespeare, a railway timetable, a common minimum programme of the government, announcements of the Nobel prizes and editorials in *Current Science* are all character strings. However, despite their widely differing contents and styles, even a casual look at them is enough to make out what they tell you.

DNA, the hereditary material, is also a collection of (simpler) character strings, made up of just the four symbols A, T, G and C. The DNA sequences also differ greatly in their 'messages'; some 'code' for proteins, others for nucleic acids, still others promote the manufacture of their own copies, while yet another set prevents the copying process. Unlike English prose, however, a casual (or even a thoroughly and painstakingly non-casual) glance is *not* enough to tell us as to which of these (or other) categories does a given DNA sequence belong. And we do need such information, the need becoming more and more pressing as the days go by. The charge of the light brigade on human and other genomes continues to generate DNA sequences at increasingly faster rates. Ironically, almost 90% of the DNA, at least in higher organisms, performs no *known* function (and is therefore crudely called 'junk DNA'). To put to use the sequence data, obtained at enormous expense, it is essential to have the means to identify whether a sequence at hand is junk or functional, and if functional, to guess its likely role.

S. Tiwari *et al.* review (**page 12**) the different methods used for making sense of DNA sequences. In principle, the approaches are no different from what we use (mostly subconsciously) to distinguish between, say, Maugham and Hemingway; most authors have a characteristic 'style' which we learn to recognize after reading them for a while. The DNA sequences (coding regions, promoters, enhancers and the like) too have their characteristic 'signatures', which can become apparent if we have a number of examples of each of the categories. However, these are not readily apparent to the human senses, (well, not quite; there are reports of DNA sequences being set to music and the connoisseurs claiming to be able to distinguish between a Beethoven-like protein region from the cacophony of a noncoding one!), and computers have to be brought in. A variety of mathematical and statistical techniques have to be used to make this process of pattern-recognition automated, objective and rigorous. Ramaswamy and colleagues describe many of these (including the ones developed by them) in detail, and also point out the criteria (sensitivity, specificity, etc.) of judging the relative merits of these approaches. Like all vibrantly active fields, there has been considerable progress in the recent past, and there is considerable scope for more improvement. This is one (and perhaps the only) area of modern molecular biology where the physicists/mathematicians/computer scientists, with their (indisputably) infinite ignorance and (arguably) unlimited intelligence, may manage to make really important contributions.

N. V. Joshi

Seeing is believing: Examining the dinucleotide frequencies in a DNA sequence

The simplest descriptors of the composition of a DNA sequence are the proportions of the four bases A, T, G and C. Logically, the next set of descriptors are the proportions of the sixteen possible dinucleotide pairs; and sequences having the same composition of the four bases can differ markedly from each other when the sixteen proportions are compared. Knowing the extent and nature of such differences is of interest to the sequence analysts. The self-evident and simplest way of doing it is to compare the sixteen proportions of the first sequence with the corresponding ones of the second – sixteen pairs of thirty-two numbers. This, however, poses a major problem. Most biologists (from the lowly taxonomists to the super-elite biotechnologists) are extremely reluctant to reduce their finding to a mere set of numbers, and even more reluctant to look at others' findings when expressed as numbers.

If all you want to do is to describe, then a picture can do the job much better – something well known for centuries. To illustrate with an unrelated example, one can more easily see the differences in the pattern of variation of rainfall throughout the year between say Calcutta and Bangalore, by looking at their mean monthly rainfall profiles. An ingen-

ious variation of this theme is adopted by A. Pan *et al.* (**page 50**). Instead of the usual *X* and *Y*-axes in a plane, they use sixteen axes, represented as arrows beginning at the origin, with an angle of 360/16 degrees between the neighbouring arrows. These axes now represent the sixteen nucleotide pairs, and the proportion (of AA, AT, . . . , etc.) found in any sequence can be marked off along the axis (arrow) corresponding to that nucleotide pair. A closed sixteen-sided polygon, formed by successively joining the points marked on adjacent axes, (or a 'map', which the authors mystifyingly call a contour diagram) shows all the sixteen proportions simultaneously. Two or more different sequences can now be compared with just a single glance at the colourful, glowing computer screen. The shapes of these polygons display the same information as contained in the sixteen pairs of numbers, but the eye, evolved over millions of years for instantly interpreting complex patterns, seems to be able to extract some meaning out of it.

It has not escaped the authors' attention that this could be generalized to examine trinucleotide frequencies and similar other features. They are also aware of the large number of different ways in which the sixteen dinucleotides could be assigned to the sixteen directions, and after some experimentation, have settled on one which gives aesthetically more appealing patterns. They do not mention a linear diagram, however, and it is entirely possible that someone else would, justifiably claiming that such a subtle change produces profound differences in the displayed patterns.

After the descriptions, follow the inferences, and the authors describe how the maps for plant, parasite and random sequences show *significantly* different dinucleotide proportions. When you have pretty pictures to convince you, why worry about such mundane technicalities as statistical tests? The appealing combination of molecular biology and computer graphics is far more persuasive than a feeble statistical phrase like 'significantly different, $p < 0.05$'. After all, it is the performance of the advertising and publicity section, and *not* statistical quality control, which make or break a company.

N. V. Joshi