Ever-longer sequences in prospect

The complete nucleotide sequence of a yeast chromosome is not only a landmark in the sequencing of entire genomes, but also a pointer to the benefits these projects will yield.

THE nucleotide sequence of chromosome 3 of the yeast *Saccharomyces cerevisiae* does not, as might be expected, appear with the article on page 38 describing how 35 laboratories in 17 countries have collaborated in its determination: there is simply too much of it. Although this is one of the smallest of the 16 chromosomes of bakers' yeast, some 45 pages of *Nature* would have been required to accommodate a simple representation (by the letters A, T, C and G) of the 315,000 or so bases in the entire chromosome. Moreover, there is little that could have been learned from the orthodox publication of these original data, important though they are. They can be stored only electronically and, likewise, searched only by machine.

That is one lesson to be learned from this venture, whose outcome is important in itself, but which is also another important landmark in the evolution of sequencing projects. In March, it was the 121,000 or so bases from three contiguous stretches of the genome of the nematode *Caenorhabditis elegans* (Sulston. J. *et al. Nature* **356**, 37–41; 1992). This month, a unique European collaboration describes a nucleotide sequence three times as long which is also, for the first time, the sequence of an entire chromosome. But these are only harbingers of a flood. The other chromosomes of bakers' yeast will punctuate the next four years, by which time the nucleotide sequence of *C. elegans* should be complete, as should be the circular genome of *Escherichia coli*. And these, of course, are only preliminaries for the various human genome projects.

That is why many readers of these documents will be concerned not only with their content but with what they promise for the future. Three questions arise, one of which is signalled by the number (147) of the authors of the article on the yeast sequence. Most of the sequencing has actually been done 'by hand', using the resources already available at the 35 collaborating laboratories, although automatic sequencing machines will no doubt be more in evidence when the project moves on to the other chromosomes. (The appearance of this article may even help to unlock the necessary equipment grants.)

The division of labour between so many laboratories is, in the circumstances, sheer necessity, given the length of the chromosome sequence, but also has several qualitative benefits. Most practically, these projects evidently need a great diversity of skills. Sequencing technique may be the bread and butter, but so increasingly is computer skill. There may be off-the-shelf programs for compiling and interrogating databases, but a feel for their advantages and limitations somewhere in the team is crucial.

Then readers of the yeast article, as of the March article on *C. elegans*, are certain to be impressed at the degree to which sequencing strategy must be changed to suit the circumstances, perhaps an awkward cluster of repeating elements or a region in which ambiguities must be resolved by cloning a specially chosen segment of the genome and sequencing that. To what extent will it be possible to substitute artificial intelligence (in the sense of AI) for these intellectual functions? Only time will tell.

There seem also to be psychological benefits in collaborations on sequencing projects. One member of the *C. elegans* team, which is a collaboration between the Medical Research Council Laboratory of Molecular Biology at Cambridge, England, and Washington University at St Louis, Missouri, calls the phenomenon "hybrid vigour". The yeast project, supported by the European Commission's Biotechnology Division, has by contrast brought together a great diversity of laboratories, not all of them equally renowned; the result seems to have been that the less well-known laboratories have been challenged to function well, while others have learned unfamiliar managerial skills.

So what do the results so far suggest will flow from the larger sequencing projects now, or soon to be, under way? The most obvious surprise, with the yeast chromosome and *C. elegans*, is that there are more genes in the regions sequenced than had been expected. The nematode, for example, is now expected eventually to yield 15,000 genes (an estimate confirmed by the same collaboration's cDNA analysis in the May issue of *Nature Genetics* (Waterson, R *et al.* **1**, 114–123, 1992)).

Chromosome 3 of yeast is also replete with previously unknown genes, or at least with open reading frames from which mRNA molecules can be transcribed when there are appropriately placed regulatory elements to turn them on. Indeed, the genetic map of the chromosome contained only 34 genes before the complete sequence was assembled, but now there are 182, not counting tRNA genes, transposable elements and genes with fewer than 300 bases (of which there appear to be only four). Allowing for a small handful of genes previously sequenced but not mapped, the result is that chromosome 3 has 145 new genes, or that its gene content is five times greater than was previously believed.

There are also, of course, the familiar oddities of improbable similarities with genes first recognized in other organisms — the *nifS* gene of nitrogen-fixing bacteria, for example. The authors themselves ask why there should be the equivalent of an essential component of the mechanism for fixing nitrogen in an organism such as yeast that does not fix nitrogen. The answer says much about the way in which chance homologies may fortuitously be invested with an air of magic. If, as is now suggested, the function of the protein product of this *nif* gene is tied up with mitochondrial metabolism, it is not all that surprising that both yeast and bacterial cells should use something very similar. The interesting question is simply that of how two such different organisms acquired such similar genes.

That question points to the third issue raised by the appearance of these articles (and the flood that will soon follow): is there not a quicker way to the listing of the genes in an entire organism by means of cDNA analysis? Dr Craig Venter at the US National Institutes of Health (NIH) has recently made most of the running in this field, which has been further dramatized by the decision of the NIH to seek patent protection for his gene fragments (which are not DNA complements of entire protein products, but simply long 'tags' corresponding to the end of them).

The truth, which the *C. elegans* and yeast studies make perfectly plain, is that cDNA analysis and the sequencing of entire genomes are not alternatives, but are rather complementary approaches. By itself, cDNA analysis can only assist with the enumeration of genes, but cannot by definition throw much light on regulatory processes, on the reasons why some genes have introns and others do not, on the intriguing tendency for tRNA genes and transposable elements to be juxtaposed and on questions such as how the genome got like that, anyway. On the other hand, cDNA analysis should prove to be a useful means of refining the definition of the probes and primers on which the sequencers have to rely. In the wake of James Watson's departure from the US Human Genome Project, there is a danger that the cDNA approach will be offered as a cheaper alternative to complete sequencing, which it is not.

John Maddox