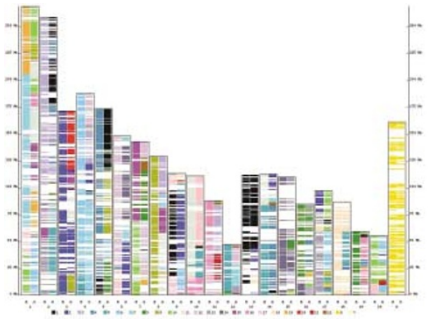
programs that are more adaptable to different users. "It is definitely not a case of one-size-fits-all," he says. "The way applications have to be developed now is to work from the consumers' needs backward." Gene-IT's new product, GenomeQuest, to be released this month, aims to provide an intranet-based sequence-search solution that makes it easy for scientists to get a complete picture of functional annotation from the entire sequence world and coordinate sequence information across diverse research teams.

There are other challenges. "The main problem, as I see it, is the pace of discovery," says Bill Ladd, senior director of analytic applications for software developer Spot fire in Somerville, Massachusetts. "Bioinformatics exists to support very dynamic, and therefore very different, claims on software development. The technology is improving all the time, and so the analysis changes. In the past, when something new came along you would have probably a few months to gather the requirements and another few months or so to roll out the new software. Now that cycle has sped up to a matter of weeks, if not days."

And the pace is only going to accelerate further. The way in which software interacts with the user has undergone a sea-change in the past few years, largely in response to the need to deal with large data sets. Ladd believes that there has been a migration towards the use of web interfaces because they are easy to manage and develop. But this has a cost. "It means that we are doing more browsing than analysis," he says. "There are databases like Ensembl, for example, that effectively collate data from different sources, but difficulties arise when you try to integrate even this collated database data with experimental data. Once I have a list of genes out of Ensembl, what happens when I ask whether other genes in my data have the same characteristics?"



Three-way synteny from Softberry.

New lamps for old

Fortunately help is at hand for the non-expert. Many of the new bioinformatics products aim to do essential tasks, such as BLAST queries, which find matching sequences in databases, and protein alignments, more efficiently and in a more user-friendly way than previous systems. New algorithms allow searches of large genomes to be done at unprecedented speeds without the loss in sensitivity that results in missed alignments. This means that it is possible to do real-time interactive searches against whole genomes and genomic data.

One such package is PatternHunter from Bioinformatics Solutions in Waterloo, Canada, which introduces the concept of 'spaced seed' to accelerate homology searches, and claims to be some 100 times faster than traditional BLAST. Whereas BLAST looks for regions where 11 consecutive residues match, PatternHunter looks for any 11 matches over, for example, an 18-residue segment, making the search more sensitive and, surprisingly, faster. Using a multiple-seed approach gives PatternHunter the sensitivity of the Smith–Waterman algorithm, but up to thousands of times faster. It runs on Sun Microsystems' Java Virtual Machine and also boasts conservative memory usage and a guarantee not to miss any alignment. The program was used by the Mouse Genome Sequencing Consortium to compare the mouse and human genomes (*Nature* **420,** 520–562; 2002), and is also used by companies such as Celera Genomics in Rockville, Maryland, and deCODE Genetics in Reykjavik.

Fast searching is also a feature of Genome Explorer from Softberry in Mount Kisco, New York, which uses the FMAP algorithm. This is a very fast algorithm developed by Softberry to map query sequence to large genomes. It keeps the oligonucleotide vocabulary of the entire genome in computer memory to speed up the searches. Softberry claims that the program can search the entire human genome for a sequence of interest in under two seconds. As well as offering simple pattern searches, retrieval of nucleotide and amino-acid sequences, Genome Explorer includes access to expression data on genes and the annotation of the draft of the human genome.

A number of desktop programs make it

SEEING IS BELIEVING

Science often works by data mining — finding correlations within and between sets of data. Dividing these sets into subsets and testing out various scenarios are key steps in this process. But many data sets generated today are extremely large and complex, sometimes involving several dimensions and varying levels of subdivision. And this is not only a concern for large pharmaceutical companies — anyone using microarrays has the same problem.

Many new bioinformatics products address this problem by organizing data in a dynamic visual context. "People are visually oriented. They are more productive when data are presented visually rather than textually," says Ron Ranauro, general manager at Paris-based Gene-IT. Data visualization aims to allow scientists to get an intuitive grasp of data structure and to spot potentially interesting trends. For example, the well known SigmaPlot package made by SPSS in Chicago, Illinois, is probably used as much for trends analysis and scenario testing as it is for the preparation of graphs for publication.

Ron Ranauro: spotting trends.

Spotfire in Somerville, Massachusetts, prides itself on the ease of use of its data visualization and decision-making software, such as DecisionSite. The functional genomics version of this program allows users to visualize genomics data and spot trends and correlations. It accepts data from a wide variety of different databases, addressing the old problem in bioinformatics that relevant data are dispersed across different locations and are often in widely divergent, and potentially incompatible, formats.

Data visualization tries to shorten the path to the 'eureka!' moment, where the researcher has intuitively grasped what the data are saying. But intuition must be backed by rigorous analysis. Programs are often packaged with a number of powerful analytical tools including similarity searches, replicate summarization and coincidence testing. DecisionSite, for example, comes with preconfigured guides to assist in common genomic analyses such as gene finding, generating "heat maps" — a type of visualization where data is colour-coded to enable an overview of large amounts of data at once. **S.B.**