Programmed for success

Advanced software and services aimed at the non-bioinformatician are making it easier to ride the tidal-wave of genomics and proteomics data. Steve Buckingham reports.

Feel that you're missing out on genomics? Even if you don't think of yourself as 'bioinformatics-enabled', your desktop PC, coupled to the Internet with its ever-expanding bioinformatics resources, will let you tap in to the data deluge. All you need to join the bioinformatics revolution is software that doesn't require you to have a deep working knowledge of bioinformatics and genetics in order to get results. But is it out there? According to software companies and service providers, the answer is a resounding 'yes'.

Finding genes lurking in raw sequence data, determining the proteins that they encode, and managing the flow of data has powered a large bioinformatics industry. Indeed, the software and services offered by big companies such as LION Bioscience in Heidelberg, Germany, coupled with specialist hardware solutions from IBM, Microsoft and Sun, have been indispensable in handling the flow of data coming out of genome-sequencing projects. But it is not just a problem of scale. As well as transforming biological knowledge, the rapid emergence of such a volume of raw data is also changing the way in which researchers frame the questions that shape research projects and the design of experiments.

But is this empowering knowledge getting to all of the researchers who would benefit from it? Biologists want data in a form that is meaningful to their own research, but they don't always have the time or resources to master the art of bioinformatics or to hire someone who has. So how can bench scientists in a small or medium-sized company or university department cash in on all of this raw information?

Help where it's needed

The need now is for software and services that can broker this information and pass it on to the non-expert in an intuitive way. "The challenge is that in-depth expertise has to be packaged into an environment that appeals to the average biologist, is easy to understand and easy to use," says Klaus May, director of sales and marketing for Genomatix Software in Munich, Germany.

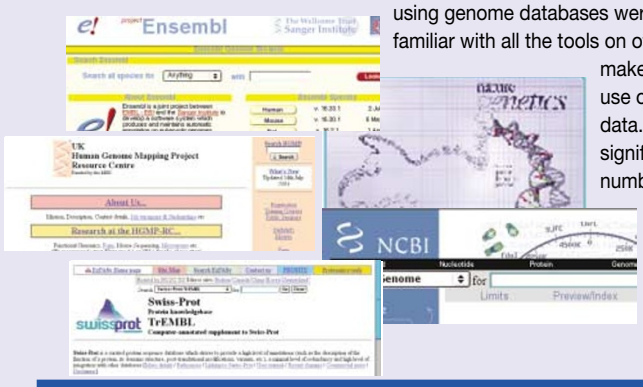
A growing bioinformatics market is trying to bridge this gap. There are encouraging signs of a trend towards software with intuitive interfaces coupled with ever-increasing analytical power. David Speechly, vice-president of the life-sciences company Applera in Norwalk, Connecticut, compares the situation to the early days of motoring. "The first cars were driven by engineers and mechanics," he says. "Later, everyone was able to drive them." Already, Speechly notes, new tools are opening up the human genome across the life sciences.

Take, for example, someone working on cardiovascular disease. They can now run a computer search of the more than 5 million known human single-nucleotide polymorphisms (SNPs, pronounced 'snips', see *Nature* **422,** 917–923; 2003) for those associated with the condition. They can then test their control groups for these SNPs to screen out those people who might have unidentified disease, thus removing one source of uncertainty from their experiment. Applera and some other companies allow academic researchers free access to their SNP database for this kind of search, and offer a catalogue of SNP-testing kits.

The bioinformatics industry itself has had to meet some tough challenges. Ron Ranauro, general manager of the Paris, France-based company Gene-IT, sees a pressing need for software firms to produce

FINDING YOUR WAY

Most genome databases now feature more than just basic sequence information; they encompass a wide range of annotated information on factors such as disease states, protein structure and polymorphisms. Yet many researchers remain unaware of the potential of these resources. A much-cited survey commissioned two years ago by the London-based Wellcome Trust revealed that only about half of the biomedical researchers using genome databases were familiar with all the tools on offer to make full use of the data. A significant number didn't even know of the existence of the European Bioinformatics Institute's public Ensembl database and website, which collates vast amounts of diverse information on the human genome. The findings prompted the trust to launch an educational initiative in 2001 to draw attention to these databases.



Genome gateways: but how to get into the garden?

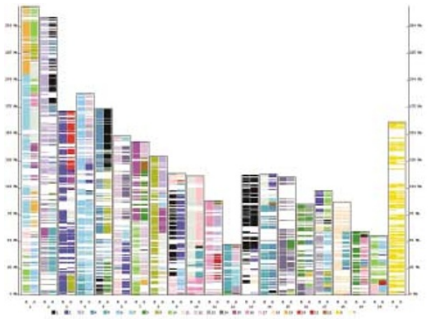
So where do you turn to make sense of the number of databases available and the array of tools they offer? A good starting point is the annual database special issue of *Nucleic Acids Research* (**31;** 1–516; 2003). The online version of this issue contains an index of major databases along with a brief summary of what each offers, information that is compiled by Andreas Baxevanis of the genome-technology branch at the US National Human Genome Research Institute.

Newcomers are often overwhelmed by the tools on offer and the knowledge of genetics that is presupposed on the part of the user. The tools tend to assume an understanding of terms and principles that are unfamiliar to, say, physiologists and pharmacologists, and without this knowledge the users cannot feel confidence in the results that they get. A useful guide for the perplexed is *A User's Guide to the Human Genome*, a web special published by *Nature Genetics* (www.nature.com/ng). This is designed for the genome neophyte, and guides the user through worked examples to give them confidence in understanding concepts and strategies that can then be used to empower their own research. **S.B.**

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.

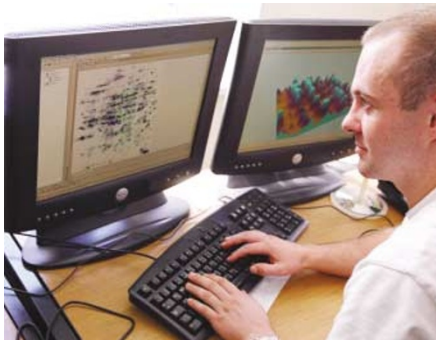
Spotfire in Somerville, Massachusetts, prides itself on the ease of use of its data visualization and decision-making software, such as DecisionSite. The functional genomics version of this program allows users to visualize genomics data and spot trends and correlations. It accepts data from a wide variety 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.**

NONLINEAR DYNAMICS

easy to do protein and nucleic-acid alignments, as well as to design primers, to search motifs and to perform multiple sequence analysis. One example is MacVector, for the Apple Macintosh and the Windows version DS Gene, produced by Accelrys in San Diego, California. MacVector has been available for many years and is continually being improved. The reasonably priced Jellyfish from LabVelocity in San Francisco, California, is also available for both Mac and PC. Like MacVector, Jellyfish will generate primers, oligonucleotides, cloning constructs and restriction maps, as well as perform BLAST searches and sequence alignments. When sequence data are downloaded, so are the annotations. Another option is the Visual Cloning 3 package from Redasoft in Bradford, Ontario, which the company claims simplifies sequence analysis, sequence editing and presentation, as well as offering access to online tools through its integrated web-browser interface and an array of plain-language 'wizards' to guide the user through the process.

Proteomics has benefited from software that not only allows rapid analysis of two-dimensional electrophoresis gels, but also draws together other functions, such as data mining, into one easy-to-use package. Progenesis from Nonlinear Dynamics in Newcastle upon Tyne, UK, comes in three modules. The first allows data annotation, the second does the image analysis and data logging, and the third is the data-mining component in which, for example, the data from different gels can be compared. Users can generate pick-lists of interesting spots



Making sense of 2D gels.

from the data-mining results and send them back to the image-analysis module to drive spot-picking from the original gel.

Paddy Lavery, Nonlinear's bioinformatics marketing manager, says that a future version of the software will allow users to import raw mass-spectrometry traces or peptide map lists into the program and to search against internal or external peptide sequence databases. The user will be able to store the search results and link them back to the gel and the sample ID of any spot in the pick list.

Making predictions

As the number of sequences of known function increase, predictions of genes, protein structure and protein function from sequence information are getting more accurate. There is a trend towards collections of integrated, coordinated suites of gene-prediction programs, many of which can be tried out on the web. Softberry, for example, offers a number of gene-prediction programs that can be accessed over the web, including FGENESH, which the manufacturers claim is fast, sensitive and accurate. The suite also includes FGENESH_C, which searches for similar cDNAs, and FGENESH+, which will find similar proteins. The fully automated FGENESH++C will automatically annotate any genome (other than human) to a standard claimed to be indistinguishable from manual annotation, using a battery of complementary techniques.

The accuracy of programs that predict protein structure from sequence has improved over the past few years and these are gradually becoming more widely used. A number of easy-to-use academic and commercial protein-structure prediction programs are available. PROSPECT Pro from Bioinformatics Solutions uses the 'threading' method, which threads the query sequence onto all known protein folds from the Protein Data Bank to find the one that gives the best-fitting structure. The program builds on this established strategy by allowing the user to feed in experimental data, such as constraints on the threading, and it checks its own results using a neural network.

Predicting the localization of a protein within the cell is also a help in identifying the function of a gene product. Softberry's ProtComp illustrates the trend in bioinformatics software to integrate diverse computational approaches. The program uses clues in a protein's sequence to guess at where it is localized within a cell, mostly by using neural networks to check sequence elements for tell-tale localization-specific motifs.

Another important issue is monitoring and controlling the flow of data as they are

GENOMIC MERGERS

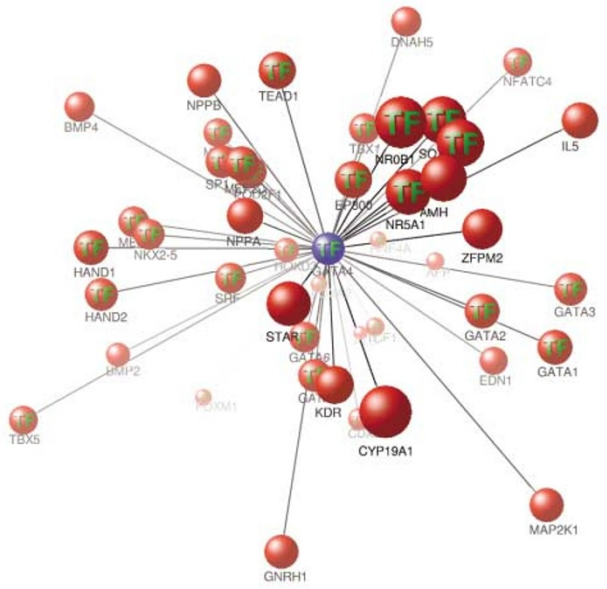
According to Celera Genomics in Rockville, Maryland, the next step after the sequencing of the human genome will be 'merging technologies'. Celera, recently forged agreements with Applied Biosystems in Foster City, California, under the umbrella of the Applera Corporation to work towards integrating all aspects of genomics. The companies are now developing an array of predesigned and prevalidated assays for genes identified on completion of the sequencing of the human genome. There are already assays for more than 18,500 human genes known to be expressed, as well as kits for over 125,000 single-nucleotide polymorphisms (SNPs) — their goal is to have 200,000 SNP assays by the end of the year.

"The most common complaint we hear from scientists is: 'but we're not bioinformaticians'," says Ramin Cyrus, a senior director for Celera. Anthony Kerlevage, a senior director at Celera, likens the situation to modern word processors. "They are packed with features, but most of us just use them to write letters," he smiles. The solution? This month, Celera and Applied Biosystems launched the 'mySCience' portal.This website allows users to upload data from instruments, databases or laboratory information management systems (LIMS), and store them in personalized workspaces. Users can then analyse their data with an array of tools, guided by predesigned workflows. These workflows act like 'wizards' — guiding the user through the options of which web-based tools are available. The site itself is free (Applied Biosystems hopes that it will tempt users to purchase the company's assays), and is designed to complement the subscription-based Celera Discovery System, which offers deeper analyses. **S.B.**

APPLERA CORPORATION

Which gene? Assays for human genetics are big business.

generated and passed through the laboratory, especially as individual experiments can now produce large, complex data sets that need to be interpreted and passed on to the next set of experiments. This is where LIMS come in — laboratory information management systems that not only store data, but will track what reagents are used and where they are bought and stored, as well as following the project's progress. LIMS have been described as electronic lab notebooks — a good LIMS package will catalogue experimental data, which it can capture directly from laboratory instruments, scientists' reports and even purchasing. Another essential feature is traceability — original data must be open to tracking and retrieval throughout the project.



Joining the dots: BiblioSphere from Genomatix searches the literature for co-citations of genes.

John Helfrich, programme manager for drug discovery and development at NuGenesis Technologies in Westborough, Massachusetts, has been keeping track of how LIMS are evolving. "I am seeing a trend towards the development of 'purpose-built' LIMS that perform a specific subset of data management within specific departments in the biopharmaceutical industry," he says. He points to the Watson LIMS from InnaPhase in Philadelphia, Pennsylvania, which was specifically designed for preclinical bioanalysis in drug discovery.

Another likely trend will see LIMS packages become easier to configure — traditionally a LIMS has been designed in consultation with the client to meet a specific need, but Helfrich thinks that future LIMS vendors will aim at more customizable off-the-shelf products. The latest version of NuGenesis' interoperable SDMS (Scientific Data Management System) was launched this June, with enhanced support for the Macintosh OS X, UNIX and Windows platforms. The system catalogues and captures all of a project's data from its source, thus avoiding the problem of traditional LIMS designed to capture only a narrowly limited data stream or restricted to using treated data. It can be integrated with an existing LIMS or any high-order IT system, or implemented as a stand-alone system for the small or medium-sized lab.

Spending less time in the library

The explosive increase in sequence data is almost matched by the increase in text publications, and keeping up with the published literature can be a full-time task. Some companies are now producing software that allows the user to explore textual data at a level beyond a simple literature search. PubGene, based in Oslo, Norway, offers a program at both free and proprietary levels that identifies potential associations of genes and proteins by finding their co-occurrence in abstracts of published papers or in gene-expression experiments. The publicly available version allows searches for co-occurrences of genes, although the database behind the commercial version is claimed to be more up-to-date and also offers protein searches.

A suite of programs from Genomatix Software takes a similar approach. One powerful member of this suite, which illustrates the emerging emphasis on the visual presentation of complex data, is BiblioSphere. This package brings together genome analysis and the US National Center for Biotechnology Information's PubMed database. Like PubGene, BiblioSphere is a data-mining tool that looks for co-citations of two or more genes of interest in published abstracts. The lowest-level search tags the citation of two genes in the same abstract. The most discerning search looks for the co-citation of the two genes in the same sentence, coordinated by a key word such as 'regulates'. The findings are presented graphically — the gene you searched for is presented visually at the centre of a sphere of related genes. Clicking on the line joining two genes leads you to the citation mentioning the two genes. Clicking on any gene leads you to further data on that gene.

Some programs even claim to be able to track that most complex domain of biological information — the research paper — to find exactly the papers you need. Despite the obvious difficulties faced by a machine attempting to 'understand' natural, human language, some packages claim to do just that. The LexiQuest Knowledge Management Suite from SPSS in Chicago, Illinois, is based on 'real' linguistics: it can race through unstructured text and produce a graphical map of the main concepts in that text. SPSS claims that the software understands natural language queries and can respond to them intelligently. "Everyone is calling text 'unstructured data', but that's not quite true," says Catherine DeSesa, senior analyst at SPSS. "Text does have structure because language has structure — a very complex structure — and it is the incorporation of the knowledge of this structure that enables LexiQuest Mine to accurately extract and organize multi-word concepts without prior knowledge of the exact terms themselves." Another example is KDE TextSense from InforSense in London, UK, which contains at its core a free-text mining toolbox and can be used to turn the information into structured data tables.

New information is driving software development, but changes are also needed to the way in which data are presented. Take the database problem — a huge number of databases now exist, all of varying quality and featuring different, often incompatible, formats. Matthew Day, databases editor for London-based online publishers BioMed Central, sees the need for change. "There aren't yet public repositories for all the different sorts of information that biologists are producing. I believe that all data sets should be published as user-friendly online databases that are closely associated with journal articles and are amenable to data mining. Thus the boundaries between journals and databases becomes blurred, and a sea of data sets is created under the umbrella of the peer-review system."

In the genome age, it is easy for small laboratories to feel that they are being left behind. Large biotech companies can now achieve in an afternoon what used to take an average-sized lab months, if not years, to accomplish. But software technologies are getting better at encapsulating expertise into compact programs. These are becoming easier to use and more reliable, and the results are easier to interpret. There is a growing emphasis on automatically drawing on diverse data sources from widely divergent locations. New services and software products are drawing us closer to the day when the expertise of a professional bioinformatician can be downloaded into a desktop computer. ■

Steve Buckingham is a neurophysiologist at the Department of Molecular Biophysics, University of Oxford. He is also a freelance writer.