Bioinformatics: Bringing it all together

Forget test tubes, petri dishes and pipettes. One of the few pieces of equipment that can be honestly labelled ubiquitous in biology today is the computer. Bioinformatics — the development and application of computational tools to acquire, store, organize, archive, analyse and visualize biological data — is one of biology's fastest-growing technologies.

Biologists at the bench studying small networks of genes want user-friendly tools to analyse their results and help them to plan experiments. They need accessible interfaces that allow them to search databases, and compare their data with those of others (see 'Genome analysis at your fingertips', below).

At the other end of the spectrum, researchers analysing whole genomes, and drug-discovery companies mining the genome for drug targets, want high-throughput analysis tools to accelerate genome annotation and extract information from databases in more efficient and sophisticated ways.

And all of those involved want more integration — integration of data across the hundreds, if not thousands, of different databases, and visual integration of data to aid interpretation. "The key to bioinformatics is integration, integration, integration," says bioinformatics expert Jim Golden at Curagen spin-off 454 Corporation in Branford, Connecticut. "To answer most interesting biological problems, you need to combine data from many data sources," agrees Russ Altman, a biomedical informatics expert at Stanford University. "However, creating seamless access to multiple data sources is extremely difficult."

Standard currencies

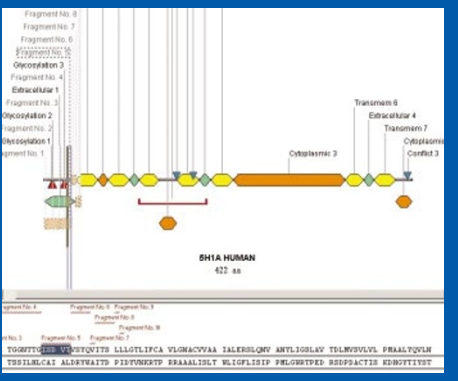
One of the most insidious problems is the lack of standard file formats and data-access methods. But attempts to standardize them are gaining momentum. One success is the distributed annotation system (DAS), a standard protocol developed by Lincoln Stein at Cold Spring Harbor Laboratory in New York and his colleagues. "It's a simple solution to a simple but obvious problem," says Stein. "There was no standard way of exchanging sequence annotations."

DAS allows one computer to contact multiple servers to retrieve and integrate dispersed genomic annotations associated with a particular sequence, such as predicted introns and exons from one server and corresponding single-nucleotide polymorphisms (SNPs) from another. It handles the annotations as elements associated with a particular stretch of genomic sequence and so enables users to obtain a picture of that genome segment with all of its associated annotations. Many providers of genome data, including WormBase, FlyBase, the Ensembl server run by the European Bioinformatics Institute (EBI) and the Sanger Institute near Cambridge, UK, and the genome browser at the University of California, Santa Cruz, are currently running DAS servers.

Reckoning that data providers will never agree on a universal standard for representing data, building database interfaces or writing access scripts, Stein thinks that web services such as DAS are the best route to interoperability. Data providers only have to agree on a small set of standards that define how their data and

GENOME ANALYSIS AT YOUR FINGERTIPS

The working biologist now has an enormous number of options when it comes to bioinformatics tools. On one hand, there is a lot of free high-quality software in the public domain. On the other, researchers can buy commercial products offering added features, such as programs to streamline sequential tasks, to access proprietary databases and to enhance data security. And because software producers realize that users' needs change and their products will rarely be used in isolation, flexibility and modularity are on the rise.



INFORMAX

InforMax's BioAnnotator uses locally stored databases to find protein motifs.

An important trend has been the increasing integration and sophistication of tools available to non-experts. A wide range of user-friendly packages incorporating tools for nucleotide and protein sequence analysis are available from companies such as MiraiBio, a Hitachi Software Engineering subsidiary based in Alameda, California; DNASTAR in Madison, Wisconsin; InforMax in Bethesda, Maryland; and Accelrys in San Diego, California. On the non-commercial side, the Biology WorkBench maintained by the Supercomputer Center at the University of California, San Diego, is particularly popular, offering more than 80 bioinformatics tools to more than 10,000 registered users. "It's a one-stop-shop for doing a lot of things," says lead developer Shankar Subramaniam. "You can be sitting in front of any type of computer; as long as you have a web browser, you can access it."

Software has also become more user-friendly. Back in the early 1990s, users of the GCG Wisconsin package, the grandfather of molecular-biology packages (now sold by Accelrys), had to work with UNIX-based systems. Although these systems are still preferred by some, users can now point-and-click their way through a wide range of tasks on ordinary desktop computers.

Another trend is the increased integration of data analysis with experimental design. The needs of bench scientists don't always coincide with those of professional bioinformaticians producing tools for whole-genome analyses. Genome projects require programs that can efficiently, if not very accurately, process huge amounts of sequence data, but the biologist in the lab is often interested in studying small sets of genes and their products with

tools are presented to the outside world.

And a 'registry' can keep track of which data sources implement which services. Scripts for retrieving a particular type of data or operation consult the registry, as they would an address book, to determine which data sources to query. A project of this type is BioMOBY, led by Mark Wilkinson at the National Research Council in Saskatoon, Canada. BioMOBY will be a powerful exploration tool, he says, because apart from answering database queries, it will discover cross-references to other relevant data and applications. Betting on BioMOBY's potential, several groups are encouraging its development. "At the moment, we have the support of almost all of the model organism databases," says Wilkinson.

Another indicator of the widespread desire for interoperability is the incorporation in February 2002 of the Interoperable Informatics Infrastructure Consortium (I3C). With 14 member organizations — including Sun Microsystems of Santa Clara, California; IBM of White Plains, New York; Millennium Pharmaceuticals and the Whitehead Institute for Biomedical Research, both in Cambridge, Massachusetts — I3C is not a standards body, but aims to develop and promote the adoption of common protocols.

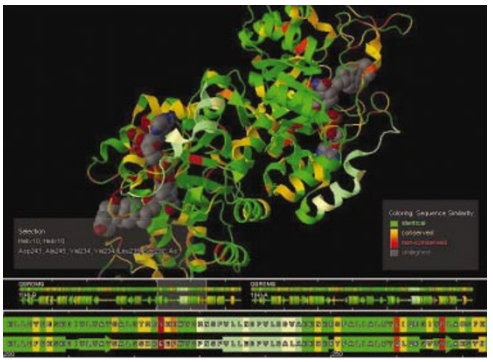
To integrate the current set of non-standardized databases, researchers are relying on two main strategies: warehousing and federation. A warehouse is a central database where data from many different sources are brought together on one physical site. Entrez, the widely used search-and-retrieval system developed by the US National Center for Biotechnology Information in Bethesda, Maryland, is an example.

Access all areas

A popular tool is SRS produced by LION Bioscience of Heidelberg, Germany, which facilitates access to a wide range of biological databases using a warehouse-like strategy. SRS is used in the online genome portals maintained by Celera Genomics in Rockland, Maryland, and Incyte Genomics in Palo Alto, California, and is the core technology of tools sold by LION.

Federation, on the other hand, links different databases so that they appear to be unified to the end-user but are not physically integrated at a common site. A query engine takes a complicated question requiring access to multiple databases and divides it into subqueries that are sent to the individual databases. The answers are then reassembled and presented to the user. Aventis Pharmaceuticals in Strasbourg, France, for example, has adopted IBM's DiscoveryLink federating software to aid collaboration between its biologists and chemists in drug development.

Which approach to use and when is much debated. "Updating and maintaining local copies of external data collections in a warehouse is a major task," says bioinformatician Rolf Apweiler at the EBI's lab in Hinxton, UK. Federation avoids this because the data are accessed directly from the original source. But the bioinformatics databases you want to query must be accessible for programmatic queries over the Internet, and most are not, says Peter Karp, director of the bioinformatics research group at the non-profit research institute SRI International in Menlo Park, California. "It's like installing a state-of-the-art telephone exchange in a village without telephones."



Structure prediction: modelling a sequence homolog in LION's SRS 3D.

LION BIOSCIENCE

Several projects combine the two approaches. On the industry side, IBM has set up a partnership with LION to

very high precision. Last month, for example, InforMax released GenomBench, a tool that allows users to predict the structure of genes and their splice variants, progressively refine these predictions, and then design experiments to validate them. "It's an interactive tool that can work with researchers not just to analyse the data they have, but to design the right experiment to resolve ambiguities in the data," says Steve Lincoln, senior vice-president of life-science informatics at the company.

Others are hooking up their software to catalogues of reagents. As just one example, the genome browser run by the University of California, Santa Cruz, is being used in a collaboration with the National Cancer Institute in Bethesda, Maryland, to identify new genes to expand, and ultimately complete, the Mammalian Gene Collection — a set of cDNA clones of expressed genes for human and mouse. The browser will be linked to the collection's website, so that users can go straight from analysing an electronic representation of a gene to ordering a clone.

A key trend in the development of commercial products is the emergence of workflows, automated chains of operations that can dramatically increase analysis throughput. For example, software producer geneticXchange of Menlo Park, California, recently demonstrated a workflow that sorts gene-expression data generated by microarrays, looks up the accession numbers that identify the selected genes, collects sequence information from the US National Center for Biotechnology Information's UniGene database, gathers annotation information from the LocusLink website, and goes to Medline to assemble a list of relevant references. "You just hit a button and it does what might take a biologist 600 hours to do, in about five hours," says Mark Haselup, chief technical officer for the company.

Some commercial products are valuable because they're linked to otherwise unavailable proprietary data. One of the main selling points of the Celera Discovery System, for example, is the access it provides to the biotech firm's high-quality human and mouse genome annotations. Unlike many other collections of annotations, a high proportion of Celera's have been generated by manual curation (see 'Putting a name on it', overleaf).

Commercial products often provide greater security for those who don't wish to manipulate their unpublished or unpatented results openly over the Internet. Although some public sites offer a degree of security, commercial packages usually have more protection options and can be operated behind a firewall.

But the recurrent theme in the design of bioinformatics tools is the trend towards increased integration. The Discovery Studio Gene package recently launched by Accelrys is a case in point. "Results are put into a project database that has the ability to be accessed by a set of applications that span both chemistry and biology," says Scott Kahn, senior vice-president of life science at Accelrys. "We set up the ability to collaborate between domains." **M.C.**

Biology WorkBench ▸ **workbench.sdsc.edu**

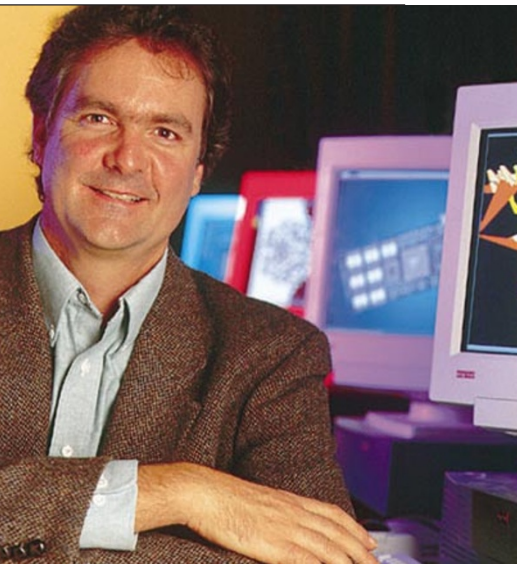
integrate DiscoveryLink with SRS. Particularly ambitious is the public-domain Integr8 project led by Apweiler. His team aims to bring together some 25 major databases spanning a broad range of molecular data, from nucleotide sequences to protein function. "We're trying to make an integrative layer on top of it all so that you can easily zoom in on the sequence data linked to the gene, and then go to the genomic data, to the transcriptional data and to the protein sequences. You'll have a sort of magnifying glass," says Apweiler.

Knowledge is power

Smart systems that can answer complicated questions about different sorts of data are also on the move. "A knowledge base is a fancy word for a database that allows you to do really sophisticated queries," says bioinformatician Mark Yandell at the University of California, Berkeley. Such databases often rely on vocabularies known as 'ontologies' (see 'Putting a name on it', below) combined with frame-based systems, a way of representing data in computers as objects within a hierarchy. One frame, for example, could be called 'protein', with slots describing its relationships to other concepts, such as 'gene name', or 'post-translational modifications'. So when a user asks a question about a protein, frames make it easy to retrieve the name of the corresponding gene and the modifications the protein can undergo. If the user asks for literature references, ontologies make it possible to retrieve not only articles that include the protein name but also those about related genes or processes.

The Genome Knowledgebase, a collaborative project between Cold Spring Harbor Laboratory, the EBI and the Gene Ontology Consortium, will have, among other capabilities, the ability to make connections between disparate genomic data from different species. "We store things specific to a species but allow a patchwork of evidence from different species to weave together," says Ewan Birney, a bioinformatician at the EBI. So when users pose questions about a biological process, they will get answers that incorporate knowledge collected from various model organisms.

Knowledge bases are being developed for a wide variety of topics, but some researchers are sceptical about their future. Information scientist Bruce Schatz of the University of Illinois at Urbana-Champaign, for example, thinks that ontologies require too much expert effort to generate and maintain. "All ontologies are eventually doomed," he says. Instead, he favours a purely automated process of knowledge generation, such as concept-switching, which relies on analysing the contextual relationships between phrases to identify underlying concepts. Concept-switching algorithms, for example, allow users to start with a general topic, such as mechanosensation, and explore its 'concept space', zeroing in on specific terms such as the mechanosensory genes of a particular species.



David Haussler: putting the picture together.

R.R. JONES

Visualizing the genome

An essential component of bioinformatics is the ability to visualize retrieved data, especially complex data, in ways that aid their interpretation. "Integration and visualization are actually very closely related, because after you integrate

PUTTING A NAME ON IT

BILL GEDDES

Lincoln Stein: bridging the gap.

A chasm separates sequence data from the biology of organisms — and genome annotation will be the bridge, says Lincoln Stein, a bioinformatics expert at Cold Spring Harbor Laboratory in New York. Spanning three main categories — nucleotide sequence, protein sequence and biological process — annotation is the task of adding layers of analysis and interpretation to the raw sequences. The layers can be generated automatically by algorithms or meticulously built up by experts in the hands-on process of manual curation.

Because manual curation is time-consuming and genome projects are generating data, and even changing data, at an extraordinary pace, there is a strong motive to shift as much of the burden as possible to automated procedures. A major task in the annotation of genomes, especially large ones, is finding the genes. There are numerous gene-prediction algorithms that combine statistical information about gene features, such as splice sites, or compare stretches of genome sequence to previously identified coding sequences, or combine both approaches. A new type of algorithm, called a dual-genome predictor, uses data from two genomes, to locate genes by identifying regions of high similarity.

Each algorithm has its strengths and limitations, working better with certain genes and genomes than with others. The GENSCAN gene-predicting algorithm, developed by Chris Burge at the Massachusetts Institute of Technology, has become a workhorse for vertebrate annotation and was one of the algorithms used in the landmark publications of the draft human genome sequence. FGENESH, produced by software firm Softberry of Mount Kisco, New York, proved particularly useful for the Syngenta-led annotation of the rice genome sequence.

Good data preparation is also important. "A lot of the magic happens in the environment, not the algorithm," says Ewan Birney a bioinformatician at the European Bioinformatics Institute (EBI) in Hinxton, near Cambridge, UK. "People often focus on the whizzy technology to the detriment of the real smarts, which happen in the sanitization of data to present them to a hard-core algorithm." Data sanitization includes steps such as masking repetitive sequences, which can interfere with an algorithm's performance.

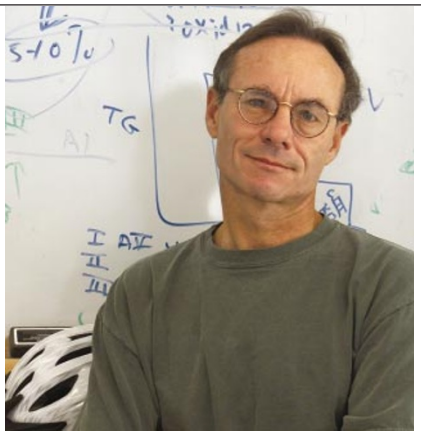
All current large-scale efforts involve a combination of automatic and manual approaches. "For me it's quite clear that they can only be complementary," says Rolf Apweiler at the EBI, who leads annotation for the major protein databases SWISS-PROT and TrEMBL. "You can't automate anything without having

information, the first thing you want to do is display it," says Altman. "They're both parts of the issue of taking information that's perfectly happy in a computer and turning it into information that a user is happy digesting cognitively."

Genome browsers are particularly powerful, as they provide a bounded framework, the genome sequence, onto which many different types of data can be mapped. The University of California, Santa Cruz, for example, maintains a browser where users can simultaneously view the locations of SNPs, predicted genes and mRNA sequences along a chosen genome stretch. "It's all about linking," says principal investigator David Haussler. "It's about having it all at your fingertips."

Tools that compare genomes from different species are also proving their worth. The VISTA project, developed and maintained by the Lawrence Berkeley National Laboratory in Berkeley, California, allows biologists to align and compare large stretches of sequence from two or more species. "It gives you a graphical output where you see peaks of conservation and valleys of lack of conservation," says Edward Rubin, one of VISTA's developers.

Spotfire of Somerville, Massachusetts, sells software that can transform all sorts of data into images. Using Spotfire's DecisionSite, researchers at Monsanto in St Louis, Missouri, represented as a 'heat map' the results of complex experiments



ROY KALTSCHMIDT/LBL

Edward Rubin takes a graphical view.

that tracked changes in the expression of thousands of genes and the concentrations of numerous metabolites during maize development. It helped them to link the expression of certain genes to the presence or absence of particular amino acids. "A lot of times it's through comparisons and comparisons and comparisons that researchers see an interesting trend," says David Butler, vice-president of product strategy at Spotfire.

Biologists are moving closer to their dream of data integration. But open issues remain. Schatz worries that if public support doesn't increase, industry may come to dominate the field, providing suboptimal solutions for scientists. "If a Celera-like company starts doing this kind of activity and they get bought by Microsoft, which is an entirely possible activity in the world at large, then it will be too late. And then scientists will get whatever the major customers of Microsoft want," he says.

But Celera's director of scientific content and analysis, Richard Mural, advocates a centralized, industry-based solution to integration and genome annotation. He notes that there are few rewards for academic researchers for working on such problems, and their focused interests can be hard to reconcile with a global approach. "To really get it done quickly and well, I think the commercial may be a stronger model," he says.

However these issues are resolved, the road ahead looks bright. "Ninety-nine percent of bioinformatics is new stuff," says Haussler. "It's an enormous frontier." ■

Marina Chicurel is a science writer based in Santa Cruz.

Distributed analysis system
biodas.org
Interoperable Informatics Infrastructure Consortium
www.i3c.org
University of California, Santa Cruz, genome browser
genome.ucsc.edu
Genome Knowledgebase
www.genomeknowledge.org
Entrez system
www.ncbi.nlm.nih.gov/Entrez
Ensembl genome browser
www.ensembl.org
VISTA
www-gsd.lbl.gov/vista

manual reference sets that you can rely on."

While Apweiler is tackling large-scale annotation, others are concentrating on finding genes and proteins linked to a particular process, such as a disease. The bioinformatics and drug-discovery company Inpharmatica in London, for example, provides annotation databases and tools to identify potential drug targets.

Because of the plethora of different names given to the same genes and proteins in different organisms, a growing trend is the use of 'ontologies' — controlled vocabularies in which descriptive terms (such as gene and protein names) and the relationships between them are consistently defined. One ontology that is now widely adopted is the Gene Ontology (GO), but it doesn't cover all biology, and others have developed their own, often complementary, ontologies. BioWisdom in Cambridge, UK, for example, sells information-retrieval and analysis tools for drug discovery based on proprietary ontologies in fields such as oncology and neuroscience.

Working as part of the Alliance for Cellular Signaling, a team led by Shankar Subramaniam is developing an ontology that captures the different states of a protein, such as phosphorylation state. This will serve as a foundation for the Molecule Pages, a literature-derived database of signalling molecules and their interactions.

GO coordinator Midori Harris at the EBI and her colleagues are encouraging developers of new ontologies to make them publicly available through GO's website. They hope this will not only drive standardization, but will help to expand GO's capabilities by allowing the creation of combinatorial terms derived from different ontologies.

But most researchers agree that tools are only part of the solution. "The passion for biology often gets missed out here," says Birney. "People think it is all about finding technical solutions that magically solve problems, but frankly, far more important is really wanting to see the data hang together." **M.C.**

Gene Ontology Consortium **www.geneontology.org**
European Bioinformatics Institute **www.ebi.ac.uk**
Alliance for Cellular Signaling **www.afcs.org**

HEIKKI LEHVASLAIHO

Automated annotation: Ewan Birney and Ensembl.