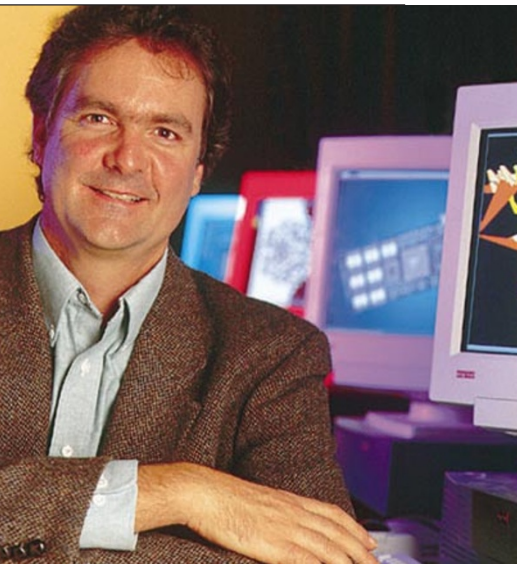
integrate DiscoveryLink with SRS. Particularly ambitious is the public-domain Integr8 project led by Apweiler. His team aims to bring together some 25 major databases spanning a broad range of molecular data, from nucleotide sequences to protein function. "We're trying to make an integrative layer on 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.

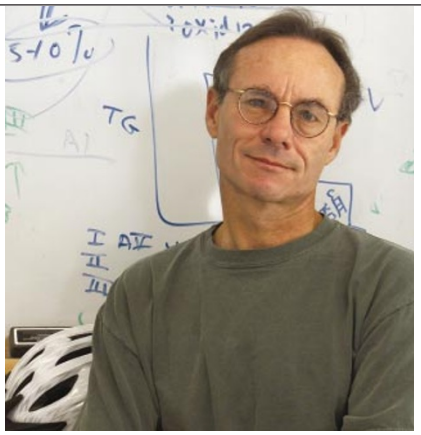
All current large-scale efforts involve a combination of automatic and manual approaches. "For me it's quite clear that they can only be complementary," says Rolf Apweiler at the EBI, who leads annotation for the major protein databases SWISS-PROT and TrEMBL. "You can't automate anything without having

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.