recognition functions, built-in BLAST search, and embedded machine-learning algorithms such as support vector machines for the analysis of microarray gene-expression data, , for example. There are also built-in routines that allow searches using 'regular expressions' — complex word-pattern matching — that complement the powers of the favourite bioinformatics programming language Perl.

LION BIOSCIENCE

Joe Donahue: different databases must work together.

But the real power of databases is the ability to unearth patterns hidden across different types of data. For this, a database must be able to query widely different types of information in a common format. "Databases are becoming more capable of doing analysis through different data types and allowing integration of different types of data," says Jacek Myczkowski, Oracle's vice-president for life sciences and data-mining technologies. For example, patterns of gene expression from patients with different forms of a disorder can be stored in a relational database table, along with written clinical notes. Algorithms such as Oracle's support vector machines can then be used to build models using these two data types to identify the gene-expression patterns that are the most reliable markers of each disease profile.

Even data mining of unstructured text has seen some astonishing advances. Oracle Text will read a document and provide an intelligent summary. "A document identified as being about cars, for example, can mention Audi and BMW and not even mention the word 'car'," says Stephens. "Oracle Text routines can extract the theme of a document like this, and can identify its subject matter".

## Working together

The integration of databases is a priority if the full potential of the genomics revolution is to be realized. "There is a clear trend today to get all these databases working together," says Joe Donahue, US president of LION Bioscience in Cambridge, Massachusetts. "Databases have always had cross-references to each other, but now we can search across them all at once."

To do this, each database needs to know something of the hidden workings of the others, such as the names of its database fields and what sort of data those fields contain. These were once closely guarded secrets, but things are changing. "The attitude only a few years ago was, 'my database is better than yours'," says Berman. "But now everyone realizes that there is far too much work to do. We have to marshal our resources."

This openness is good news, but will databases ever merge seamlessly? Myczkowski is pessimistic: "There can be no permanent standards because of the pace of change in the data." Steve Gardner at text-database company BioWisdom of Cambridge, UK, agrees. "You will never get people to adhere to standards enough to semantically integrate databases," he says. "There have been strides made in the technology to map data structures together using rule-based or ad hoc strategies, but all these systems fall down because they need rules that link fields from one database to another." But it is not all gloom. Run a query against your favourite protein at the European Bioinformatics Institute (EBI) website, and you'll see it run seamlessly against a host of diverse databases housed at separate institutions and developed by different authors with different uses in view.

## Database maintenance

For most research groups, however, setting up their own database of any significant size or complexity is not easy. Even when finished, a database needs to be updated regularly, the new data have to be parsed, indexed and stored, and special software often has to be developed. So, despite the desirability of an in-house, home-made database, the cost of maintaining it can be prohibitive for a small research group.

Paris-based Gene-IT aims to fill this gap in the market. Later this year the firm will launch its GenomeCast automatic

# GETTING THE MEANING

Although a relative newcomer to bioinformatics, ontologies have already attracted commercial interest. BioWisdom of Cambridge, UK, supplies ontologies in various fields. "Life science R&D poses a multidimensional problem," says Steve Gardner, BioWisdom's chief technical officer. "The problem is being able to communicate the information to a user interested not just in a molecule, but also in the context surrounding that molecule." BioWisdom currently offers more than 10 million distinct concepts linked by over 100 million relationships.

BIOWISDOM

BioWisdom can also assist researchers to develop their own ontology. The first task is to build a database framework to encapsulate it. An additional framework embeds methods to normalize the incoming data, so that an entity is recognized despite having different names in different data sources. This is not easy: the sedative diazepam, for example, has some 197 synonyms.

Good ontology software can even help the researcher develop new hypotheses. "We have inferencing programs that draw together different concepts," says Gardner. "If one ontology says that COX2 is expressed in synoviocytes, and another says that synoviocytes are implicated in rheumatoid arthritis, the inferencing program would suggest that COX2 may be implicated in rheumatoid arthritis."

The output of an ontology is a graph: a representation of the relationships between concepts. Once a graph has been generated, users can then bring their experience to bear. For example, they can exclude types of information on the strength of the evidence. "We call this a semantic lens," says Gardner. "You pass this lens over the data and it filters them out like a polarizing filter. This makes a new graph that lets you highlight the interactions that are interesting to you." BioWisdom's system has a hierarchical family of relationships: the protein-to-protein class, for example, has 400 potential relationships (such as 'interacts with', 'upregulates' and 'activates'). Thus, ontologies allow the user to search using one key term by resolving the meaning of that term, and then searching against it.

Steve Gardner: linking concepts.

A taste of how ontologies work is provided by the public-domain Genome Ontology (GO) Browser, which gives free access to the genome ontologies developed by the GO Consortium. Three ontologies have been developed: molecular function, biological process and a cellular component. Using the Ensembl GO browser, the user can find the Ensembl genes that have been mapped to these ontologies. The search term is presented at the centre of a 'mind map'. Clicking on a 'child' or 'parent' term will produce a new Ensembl GO report centred on that term. The genes found are listed, along with links to different types of views of each gene and its chromosomal location. The ontologies can be also searched directly, with the results showing the connections between the terms. **S.B.**