Data's future shock

Databases are having to move with the times as people expect more from them than simple data storage and retrieval. Steve Buckingham investigates.

PDB

Adding value: the PDB team now curates more than 22,000 protein structures.

Today's databases are faced with a moving target. There is so much more information available, and the type of data being collated is changing. Decisions have to be made: how should a database be restructured to accommodate the new information? What do we do with the old data? Are they compatible with the new results, or must they be hived off and archived in some way?

Take protein structure databases. Not long ago it was enough to submit a set of atomic coordinates that described a protein's structure. Now, such databases are expected to store 'meta-data' as well — how the protein was produced and purified, and how its structure was solved. And the rise in high-throughput projects will make yet greater demands.

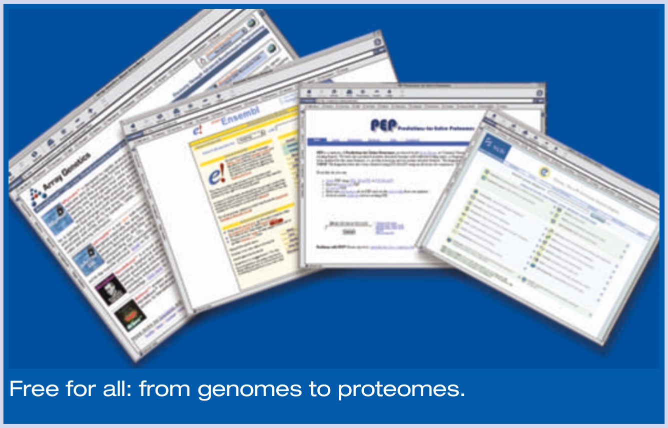
But administrators of public databases are keeping pace with the changes. When the Protein Data Bank (PDB) was set up at Brookhaven National Laboratory in 1971 it held seven structures. Today it has more than 22,000 X-ray and NMR protein and peptide structures. And whereas it was once the exclusive haunt of crystallographers, the PDB is now regularly used by biologists of all kinds.

The success of databases such as the PDB in keeping pace with these changes is due largely to careful planning. "We are preparing for three challenges for the future," says Helen Berman, the PDB's director: "The effects of structural genomics, the need for better storage of macromolecular complex data, and internationalization. There's a lot of change in the pipeline." Berman expects structural genomics to double the number of data items attached to each protein submitted to the PDB. The database is also ready for the increasing interest in macromolecular

EXPLORING THE PUBLIC DOMAIN

A vast body of annotated and linked data is available in the public databases. But how do you find the database that best fits your needs? One place to start is the supplement produced as the first issue of each year by the journal *Nucleic Acids Research*, which is free online.

Alternatively, you could plunge straight into one of the large, general-purpose, bioinformatics databases such as the Ensembl Genome Browser (EGB) or the National Center for Biotechnology Information's Entrez portal. Most are now so closely integrated with more specialized databases that navigation through a collection of databases is all but seamless. Careful design makes them powerful tools even in the hands of a novice, yet it is easy to progress to more sophisticated use. Simple search-text boxes take a term through a selection of databases, and various options control the amount and type of data returned. Help buttons, along with links that provide simple explanations of each term, are never far away.



Free for all: from genomes to proteomes.

For genomics, the EGB is a popular port of entry. This collaborative effort between the Sanger Institute, the European Bioinformatics Institute (EBI) and the European Molecular Biology Laboratory provides automated annotation of human, mouse, rat, fugu, zebrafish, mosquito, fruitfly and nematode genomes. The homepage contains a link to an 'Ensembl tour' and 'worked examples'. Users can search for a term across species for protein, disease or mRNA, or can follow links to a page dedicated to each species.

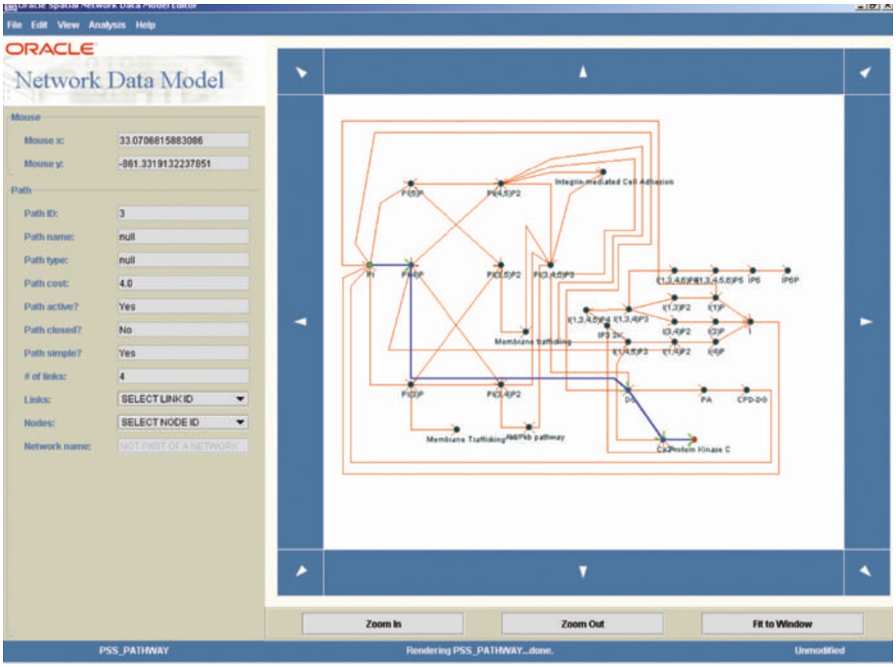
PEP, a database of Predictions for Entire Proteomes, is the result of a sophisticated analysis of proteomes from over 60 species. The work of Burkhard Rost's bioinformatics group at Columbia University, New York, PEP primarily contains open reading frames (ORFs) along with predicted structural domains detected within the ORFs. PEP can be searched online either directly or through the EBI. Alternatively, the entire PEP database can be downloaded — if you have space.

Some commercial companies offer free online access to their own high-quality databases to academic researchers, such as the MendelBase database of structural and functional protein information from Array Genetics of Newtown, Connecticut. **S.B.**

complexes, such as viral assemblies and ribosomes. Integrating these new data with the existing store is not easy but, as Berman says, "As science moves, the database must move with it." Beth Smith, director of solutions development at IBM in Somers, New York, agrees. "Annotation is going to lead to a huge increase in volume data," she says. "As medicine moves towards targeted treatment as a result of genomic approaches, we will see the rising need for high-performance computers and storage hardware. We aim to stay ahead of that capacity."

Planners have had to develop strict but extensible standards. In the case of the PDB, these took the form of the 'macromolecular dictionary' format. This has some 1,700 terms that not only define a protein's structure but also how that structure was solved. It encapsulates details of data types used in crystallographic descriptions, as well as the relationships between those data. And it is expandable — new entries are made according to strict procedures, so that new data types will always be fully integrated with older data.

As databases have changed, so has the software underpinning them. Database software company Oracle, of Redwood Shores, California, already has some 75–80% share of the general database market worldwide, and two years ago it turned its attention to the lucrative life-sciences market. "We are an opportunistic organization," says Susie Stephens, a senior life-sciences product manager at Oracle. "We see that the life-science database area is a substantial and sustainable business."



Mouse
Mouse x: 33.0706815883086
Mouse y: -861.3319132237851

Path
Path ID: 3
Path name: null
Path type: null
Path cost: 4.0
Path active? Yes
Path closed? No
Path simple? Yes
# of links: 4
Links: SELECT LINK ID
Nodes: SELECT NODE ID
Network name: NOT PART OF A NETWORK

Pathway to knowledge: Oracle's Spatial Network Data Editor.

Database software is striving to meet the demands of this market. Users want access to distributed data with full integration of different data types. Technologies embedded in Oracle's database software, for example, allow a query to be run across distributed databases of different types, including non-Oracle and flat-file databases. Users also want to manage large quantities of data, and to be able to adjust the capacity of their hardware to the size of their database and the demands placed upon it. Oracle's answer is Real Application Cluster (RAC) technology, which makes it easy to add new servers, or nodes, to an existing set of servers on the fly, in response to demand, and without having to reconfigure the whole database.

Oracle's new database release, 10g, is its first to incorporate features specifically geared to the life sciences, such as pattern-

BUYING INTO THE KNOWLEDGE GAME

Despite the impressive public databases, commercial ones can sometimes offer added value and convenience. They typically incorporate at least some information that is not available in the public domain, and have also done much of the hard work of annotating sequences and collating genomic and proteomic information.

Iconix Pharmaceuticals of Mountain View, California, for example, offers the DrugMatrix chemogenomics database and informatics system, which integrates public-domain chemical data with thousands of results from its experiments on the effects of known drugs and related compounds on gene expression and cell biology. DrugMatrix can help predict the effects of a test compound on gene expression and identify compounds that have similar effects to those in the database.

In its Discovery Knowledge database suite, MDL in San Leandro, California, offers two chemical databases, CrossFire Beilstein and CrossFire Gmelin, covering organic and inorganic chemistry, respectively. These databases are installed on a local server for access through proprietary browser software. MDL also offers Biopendium from Inpharmatica in London, which enables researchers to identify known drug targets and select related proteins in a range of experimental model systems. It uses comparisons of sequence, structure and ligand interactions, presented via a interactive alignment editor, ligand-interaction viewer and three-dimensional structure viewer. MDL's Discovery Gate structure-searchable literature information resource, combining 17 chemistry-related databases, is now also available on an academic licence.

Bringing a variety of information together in one convenient package is the selling point for commercial databases. For smaller research departments, data purchasing can fill big gaps in research capability. Buying databases can, for example, effectively bring high-throughput approaches within their reach. BioMax Informatics of Martinsried, Germany, for instance, offers reasonably priced subscription access to an annotated human genome database. The most recent release also includes the mouse genome and is integrated with the ProChart protein-interaction database from peptide-synthesis company AxCell, in Newtown, Pennsylvania.

Available online through an academic or commercial licence, the LifeSeq Foundation database from Incyte in Palo Alto, California, provides manually annotated and highly collated data on the sequence, expression and function of some 18,000 complete human genes and many more expressed sequence tags, including proprietary data not available in public databases. Each gene or gene fragment in LifeSeq Foundation is annotated with comprehensive functional information, including its relevance to disease. The database also contains information on the tissues in which a gene is expressed, related genes in the human genome, counterparts in model organisms, and known mutations. Incyte's ZooQuest database extends LifeSeq Foundation to cover mouse, rat, monkey and dog, and its Proteome Bioknowledge Library complements these databases with manually curated information gleaned from the literature on protein function and interaction for humans and selected model organisms.

S.B.

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 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.

BIOWISDOM

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.**

database-update system, complementing its GenomeQuest sequence-search system, which has recently been adopted by the European Patent Office. GenomeCast is aimed at both small labs and large drug firms. GenomeCast will automatically perform regular database upkeep without human intervention. As new data are posted on public databases, the program will aggregate them online, annotate them and combine them into a common format native to the GenomeQuest search engine. This eliminates the need for continual monitoring by a database administrator, and is rather like having someone doing your database administration for you remotely. Ron Ranauro, general manager of Gene-IT, believes that GenomeCast is following a trend in the database field. "The game has shifted away from providing curated scientific content towards delivering increasing amounts of data in real time along with the best tools at a reasonable cost and within a reasonable IT framework," he says.

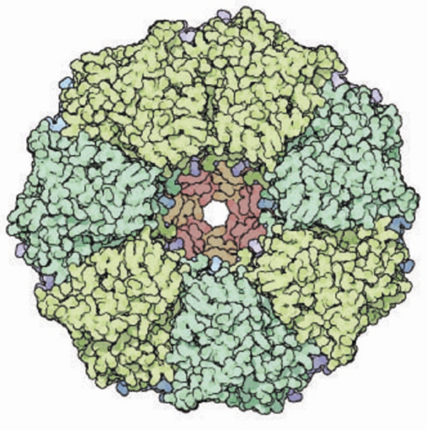
Let's talk

When it comes to databases talking to each other, there are five broad approaches to the problem of relating data entries from different databases: rules-based approaches, data warehousing, search optimizers, federation and ontologies.

Rules-based systems operate by specifying explicitly the relationship between different fields in different databases. This approach relies on records of the same object in different databases sharing some identifier, or cross-reference. With genes, for example, this might be the GenBank accession code. LION's SRS technology is the rules-based system that underlies the interoperability of the European Molecular Biology Laboratory, Wellcome and Sanger Institute databases, along with nearly 20,000 other commercial and academic databases. Users of SRS have made the details of their database structure, along with the parsers of their data, available to the SRS system. So, as more institutions use SRS, it can integrate more data, creating an ever-widening circle of interoperability.

SRS is bundled into a coordinated package, SRS Evolution, along with SRS Relational (for accessing relational structures), SRS 3D (for integrating protein structures), and other modules that assist data downloading and expression analysis. SRS technology also underlies LION's DiscoveryCenter software, aimed principally at the drug-discovery market, a package that allows a single point of access to a number of databases with integrated analysis applications.

LION's collaborations with the pharmaceutical industry are resulting in new software solutions. "We are expecting a big surge in the field of pharmacogenomics," says Simon Baulah of LION. "Our customers are starting to use SRS to integrate patient data and gene-expression data in improving personalized medicine." A collaboration between LION and the Cambridge-based UK Human Genome Mapping Project has resulted in integration into SRS of the recently developed EMBOSS query suite, a free set of bioinformatics applications that rivals the performance of the commercial Wisconsin package from Accelrys of San Diego, California.



What's in a name? This chaperonin complex can be tracked through the databases by its PDB ID 1aon.

D. GOODSELL/SCRIPPS RES. INST.

An alternative approach to database unification is warehousing — making a local copy of data drawn from diverse sources and then forging them into a common format in a unified, specialized database. This approach can be expensive in time and money, but commercial warehousing solutions from companies such as Iconix Pharmaceuticals in Mountain View, California, Incyte in Palo Alto, California, and Gene-IT could be one answer. Alternatively, tools such as DS SeqStore from Accelrys make in-house data warehousing easier. Using a client/server architecture built on Oracle, this program helps to set up a secure database complete with analysis tools and a complete GenBank, GenPept, SWISS-PROT with SP-TrEMBL distribution. Its open architecture makes it comparatively easy to adapt the design to the user's corporate or lab needs.

Another way to search across different databases is to use query-optimizing systems. These use a battery of strategies to recast the query until the best results are returned from the databases. This is the approach behind the Discoverylink system from IBM, which uses a set of 'wrappers' to adapt a query to the databases questioned. "Discoverylink allows optimization of searches across a number of different databases with diverse formats, but the user is presented with only one interface," says Smith.

In the federation approach, member databases agree to represent data in a certain way, so that no adjustment has to be made to harmonize them. An example is FEDORA (Federation of Research Assets), a federation of six special HTTP servers, released by Metaphorics of La Jolla, California. Data are federated into a knowledge base comprising a set of hyperlinks between synonyms and near synonyms, which permits sophisticated data mining. The current FEDORA cluster includes Empath, a server for metabolic pathway information, Planet, a server for protein–ligand information and WDI, a server for the World Drug Index.

An understanding approach

Human languages are rich in synonyms and subtleties of vocabulary. But this very richness makes cross-database searching a hit-and-miss affair. This looks set to change with the introduction of new techniques, such as ontologies, that are beginning to grapple directly with semantic complexity. Ontologies are networks of objects, their properties, and their relations to one another. They try to tackle the meanings of words, rather than just treating them as strings. This means more than just reducing linguistic complexity: ontologies actively exploit it.

Ontologies allow a concept in one resource (or database) to be mapped onto a concept from another. For example, an ontology might contain a representation of the fact that muscarinic acetylcholine receptors are G-protein-coupled receptors. This would allow a search for G-protein-coupled receptors across different databases, even though the search term 'G-protein-coupled receptor' might not occur in all of them.

In an ontology, the core concept (a gene for example) at the centre is connected to related concepts (such as coexpressed genes, proteins, diseases, tissues or compounds) which in turn are connected to yet more concepts (see 'Getting the meaning', opposite). But the links in ontologies are much more versatile than just a simple line, they can express relationships such as 'BINDS-TO' or 'IS-EXPRESSED-IN'. Ontologies are dynamic maps of information space and, like sourdough, once you have got it started, you can go on adding to it.

Ontologies are still in their infancy. But if they deliver what they promise, their contribution to making new insights could be enormous. ■

Steve Buckingham is a neuroscientist at the Medical Research Council's Functional Genomics Unit at Oxford University, UK.

PDB
- **www.rcsb.org/pdb**

Ensembl GO browser
- **www.ensembl.org/Homo_sapiens/goview**

Nucleic Acids Research database issue 2004
- **nar.oupjournals.org/content/vol32/suppl_1**

Predictions for Entire Proteomes
- **cubic.bioc.columbia.edu/pep**