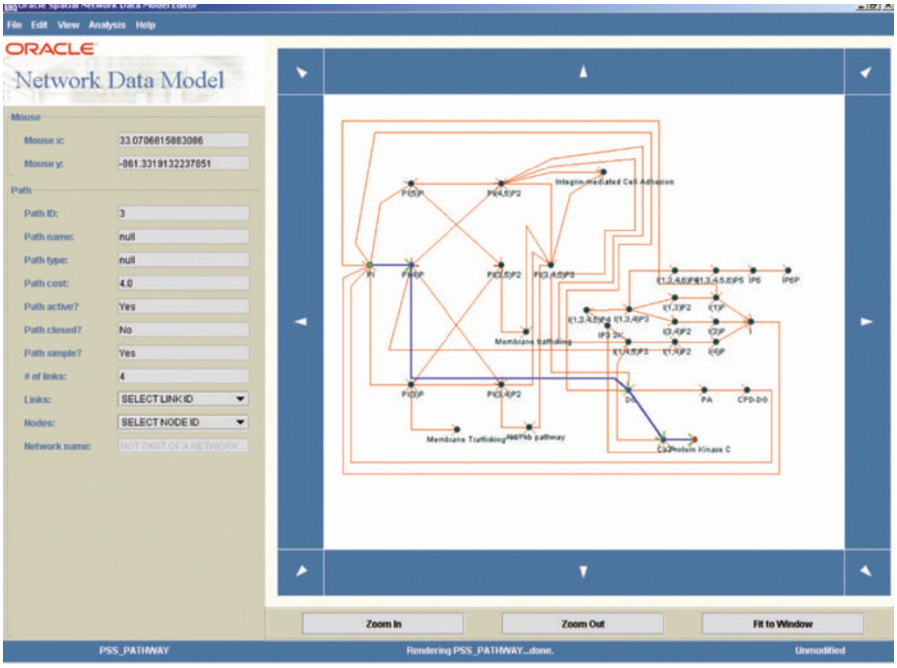
complexes, such as viral assemblies and ribosomes. Integrating these new data with the existing store is not easy but, as Berman says, "As science moves, the database must move with it." Beth Smith, director of solutions development at IBM in Somers, New York, agrees. "Annotation is going to lead to 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."



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.**