nature structural & molecular biology

Data, data everywhere…

© 2005 Nature Publishing Group http://www.nature.com/nsmb

Data: we generate it, publish it, share it and store it. Nowadays the sharing and storing involves an extensive array of databases. No doubt you've noticed that in addition to the Protein Data Bank (PDB) codes at the ends of many of our articles, some now have accession numbers from the world's largest repository of protein interaction data, the Biomolecular Interaction Network Database (BIND, http://www.bind.ca). No one can argue with the usefulness of these databases to the structural and molecular biologists who ultimately want to understand the relationship between a gene product and a phenotype. However, collecting and comparing data is an overwhelming task for the time-pressed scientist. In theory, one search engine could query all databases and obtain information on the gene as well as all relevant interactions, pathways and structures, but in practice there are both legal and technical impediments to cross-database communication.

Web search technology, which is exceptional at finding the location of specific types of data, has not lent itself well to effective querying across multiple databases. In part, this is due to the lack of intellectual property models that cover databases. Protection of deposited data now consists of digital locks that prevent or limit access to the contents. Furthermore, firewalls and similar access restrictions put in place to protect the repository from malicious corruption by viruses and hackers also impede communication.

Other technical issues have also prevented effective cross-database communication. These include the inability of search engines to recognize the data in a related repository and integrate it efficiently with information from other sources. For example, many databases, such as BIND and company and academic websites that collect information on protein interactions, do not share a compatible data structure that would allow effective cross-database communication. Search engines are unable to recognize the format of data or may not be able to incorporate them into an existing repository because data are not categorized using a common vocabulary. However, BIND and many other databases are now supporting a programming language and vocabulary proposed by the Human Proteome Organization and members of the Proteomics Standard Initiative (http://psidev.sf.net). These definitions of common data formats should help promote the exchange of information between diverse groups of data producers. Now it's simply a matter of getting database curators to use one standard language and vocabulary.

Scientists are also hard at work developing user-friendly interfaces in which data related to a query are presented in a standardized and digestible format for users. Jargon is limited and key data are highlighted. For example, Biozon (http://www.biozon.org) compiles the holdings of more than 12 databases containing DNA and protein sequence information, protein structure (PDB, SCOP), gene expression and ontology, metabolic pathways (KEGG) and protein-protein interactions (BIND). Furthermore, Biozon allows searches across different data types, facilitating cross-database comparison and the identification of discrepancies or inconsistencies between datasets.

User-friendly interfaces are a step forward, but what ensures that the myriad databases will still be there to query five or ten years down the line? The future is uncertain for many reasons, including a lack of time on the part of the developer to continually curate the database, manually ensuring that only high-quality data is added to the repository, and to update the computer hardware to ensure ready access. Even simple lab databases can quickly grow into unwieldy beasts that require more time from a scientist who has experiments to do. Curators and hardware require a continuous stream of funding, so loss of financial support is perhaps the more significant threat to the survival of a database. The financial resources for many databases come from the public sector and are always in danger of disappearing. For example, Biozon is funded by the U.S. National Science Foundation, but it is the additional funding from Sun Microsystems and Cornell University that may ensure its longevity. BIND, which has been funded by the Canadian government since its inception in 1999, has been seeking new financial support since the Canadian agencies failed to renew its grants. BIND is likely to receive financial support from Singapore, which has provided some support in the past. As many of us know, grant support can come and go, but measures can be taken to ensure that future funding will be garnered.

So what characteristics are important for a database to survive changes in funding? Successful databases such as the PDB have continued to improve their user resources, whether it is their search engines, user-friendly interfaces or available data analysis and visualization tools for specialists and nonspecialists who need to use their data. More recently, these established databases have begun to link to other successful databases. For example, the PDB is integrating its content with data from Gene Ontology, Enzyme Commission, Kyoto Encyclopedia of Genes and Genomes and various National Center for Biotechnology Information resources. These data are mapped onto structures and loaded into the PDB database. BIND has a similar process by which they integrate the PDB data into their repository. Such integration and implementation of user-friendly tools entrenches databases in the daily lives of scientists and makes it easier for repositories to convince funding agencies that investment in their databases is worthwhile.

Although not all databases will survive the technical and financial pressures they encounter, structural and molecular biologists can rest assured that using databases will become easier in the near future. The current trend toward designing better web interfaces that collect information from multiple databases will facilitate access to diverse data and an increasingly comprehensive understanding of cellular systems. ■