Search and retrieve

Large-scale data generation is becoming increasingly important in biological research. But how good are the tools to make sense of the data? • *by Alfonso Valencia*

Biological research has produced a large amount of useful data that has only been obtained through significant effort. A good publication in a peer-reviewed journal takes, on average, 2–3 years of laborious and time-consuming experimentation, often involving more than one

the underlying data and storage formats that make it difficult for scientists to retrieve and analyse the information that he or she needs.

The initial demands of molecular biology and biochemistry to amass information about genes, proteins and

All major biological databases are now implemented internally as relational databases that store data in tables, index them at various levels and define connections in order to facilitate the formulation of specific queries. For example, Genequiz (Andrade *et al.*, 1999), a system for automatic genome analysis developed in the early 1990s, stores every result of a sequence search by linking proteins similar to the query protein in one column, the corresponding similarity values in another column and the length of the alignments in a third column, all interconnected by the number of the gene sequence stored. This scheme is very simple, since it includes only links between sequences, similarity values and sequence lengths and enables elementary similarity searches to be carried out by sorting the corresponding columns and displaying the information associated by their position in the same row.

A more sophisticated model involves the construction of more complex relationships, which requires links between information items that are not the simple 'be in the same row of the table' relations of the previous example. Such a model is called an 'entity-relation scheme', and this forms the core of a database, since only what is initially separated (entities) and linked (relations) can later be retrieved. The language associated with

The limitations of the tools and systems to store, sort, interlink and order the >500 data resources are becoming increasingly apparent

researcher plus a number of technicians in order to produce only 6–10 pages of new information and possibly a few words in a specialised database. But new technologies and increasing automation, such as the use of robots and massive parallelisation, are now in place for generating huge amounts of data, whose proportions are perhaps only comparable to that generated in astronomy and nuclear physics. Systematic harvesting of this data would considerably ease research in molecular biology, where a major challenge will be to describe complete biological systems using integrated approaches—the 'system biology' paradigm.

To fully capitalise on this promising scenario, the data must be accessible in a rational and easy way. Pure computation and speed of data transmission are constantly improving and represent less of a problem than the development of the corresponding software, which has become the real bottleneck. The limitations of the tools and systems to store, sort, interlink and order the >500 data resources, ranging from sequence to pathology and gene expression to metabolism data are becoming increasingly apparent (for a collection of data-bases, see <http://www.infobiogen.fr/services/dbcat/>). The problem is further exacerbated by the large heterogeneity of

protein structures were satisfied with simple text-based repositories. The sequence, possibly the name of a protein, its origin and a brief description of its putative functions were the basic content of early databases such as EMBL (Stoesser *et al.*, 2001) and PDB (Berman *et al.*, 2000). This storage model was very useful for small amounts of data and the computational limitations at that time, but it had several shortcomings. First, it was not sufficiently systematic, since it lacked a set of well-defined functions with unique identifiers. Secondly, it was not useful for finding correlations between data. Thirdly, it was not adequate for the formulation of questions (e.g. to find all ATPases with transmembrane segments). To perform such queries, external programs were required to scan the information for keywords. But this cannot be achieved if information on function

The challenges of organising, combining and interpreting biological information are more complex than in many other scientific areas

(ATPase) and structure (transmembrane) are not stored in well-defined separate fields that can be accessed directly. It became obvious that this text-based format would not be sufficient to retrieve all forms of useful information from the data.

the database, generically known as SQL (Structured Query Language), is especially helpful in that it allows the formulation of more complex questions that combine values from distinct columns (e.g. limiting the alignment length for a given detected homology). An example of a relational

Fig. 1. Example of information contained in a database and the formulation of the associated queries. This case corresponds to the fragment of an entity-relationship schema necessary to answer the question: ‘What are the species, and their taxonomy, found in a BLAST search?’ The sequences are linked to chromosomes and from there to genomes. The genomes correspond to a given type of organism for which taxonomical classes are available. At the same time, the sequences are linked to proteins that are used for the BLAST search. For the system to work, the entities (organisms, genomes, etc.) have to be defined accurately, as well as the connections between them (‘is a’, ‘belong to’, ‘contains’, etc.). In this schema, the database managing system enables the formulation of queries and the retrieval of the corresponding relations used to answer the query. In the inset, the explicit SQL query necessary to answer the question is displayed. The query requires the detailed knowledge of the schema (entities and relations) and is formulated in a particular language to enable the operations of retrieval and comparison.

database schema and the associated SQL query is shown in Figure 1. The Regia database (Alonso-Allende *et al.*, 2002) may serve as an example of a complex relational schema that includes data about protein interactions, gene expression, metabolic profiles and phenotypes of a large collection of *Arabidopsis* transcription factors and their mutants. Such an organisation obviously comprises many more entities, with numerous connections at various levels. To fill the database tables with the corresponding data, to ensure the consistency between their relations and to allow the formulation of queries, the biotech and pharmaceutical companies usually use commercial database managing systems such as Oracle (Redwood Shores, CA) and Sybase (Dublin, CA), whereas the academic sector uses open managing systems such as PostgreSQL (<http://www.ca.postgresql.org/index.html>) and MySQL (<http://www.mysql.com/>).

Further developments in database standards are influenced by new methods of representing information, especially

the HTML language used to describe web pages. HTML has had a great impact, and its current evolution, XML (eXtended Markup Language), has yielded an attractive form for encapsulating and sharing information that goes far beyond the representation of web pages. Numerous initiatives are now under way to define XML standards in various domains (Achard *et al.*, 2001), including software

that facilitate the manipulation and organisation of objects and relations in terms of classes and shared properties (inheritance relationships). A recent example of an object-oriented database is the ArrayExpress system developed by the EBI (<http://www.ebi.ac.uk/arrayexpress/>), which is internally based on an object-oriented schema. For various technical reasons, object-oriented databases are still less common in molecular biology than in other domains.

Even if all the basic database technology described above has been created and implemented in other domains and areas of technology, the field of biology presents its own interesting difficulties and peculiarities. The rapid evolution of technology with the continuous emergence of new data types, the strong interdependence of various types of experiments, the diversity of biological systems with intrinsic specific features, their strong interconnection, and the continuous requirement of additional analysis external to the experimental data—such as sequence searches, alignments, protein models—makes the challenges of organising, combining and interpreting biological information more complex than in many other scientific areas.

Equally essential for the success of a database is the quality of the stored information. The challenging task for database managers is to coordinate the annotations of entries with features extracted from the literature, with the skilful use of analytical methods and, crucially, the maintenance of previous annotations. Owing to the complexity of these operations, early annotations were not properly linked to the supporting evidence. For example, the assignment of a protein function based on sequence similarities was not linked to

The common practice of comparing sequences and assigning the same function to similar novel proteins produces a considerable number of errors, since sequence similarity does not necessarily imply functional similarity

developments that directly include XML-encoded information in the appropriate database architecture.

The complex, large data structures that can be contained in relational databases are difficult to create and maintain. The object-oriented technology addresses this problem by creating database structures

the corresponding alignments, scores and evidence used in the analysis, which hinders subsequent testing of the annotations after a few years of continuous input of new information. A recent analysis (Blaschke and Valencia, 2001) of a modern protein interaction database revealed that only ~30% of the information

Terminology in databases

Flat file. A common standard text file, like those generated by simple edition programs. Data repositories constructed as flat files require long searching procedures, since every character has to be scanned to answer every query.

Database. A repository of information organised under a logical structure, including fields and relations. The term is used in molecular biology to refer to any collection of data, even if they are only stored in a flat-file format, with a minimal structure.

Pointers/links. The relations between entities (sequences) in different databases. In some systems (e.g. SRS), these links are stored specifically in separate files, which can be searched and retrieved by the system.

Relational databases. An organised collection of data and relations, usually constructed under a given entity-relationship schema. Different systems are available for the creation and manipulation of relations databases (database management systems) in what is the more extended type of database structure.

Entity-relationship schema. A well-defined set of entities connected by different types of relations. Deriving a good schema requires considerable effort, a deep knowledge of the field and a certain capacity for predicting the key questions formulated by future users. This initial step in the design of a database is vital for its future and normally involves the participation of software engineers and domain experts. Their decisions about what part of reality to introduce in the database are very difficult to reverse once the database is implemented and filled with data.

Queries. A question formulated to a database. In the common case of a relation database, the queries are formulated in a specific language called SQL.

SQL. The general language used for formulating questions to a database. It includes the possibility of formulating logical operations. A number of systems include graphical interfaces to simplify the use of the SQL language.

Object-oriented databases. Databases representing a more sophisticated database concept that implies more general relations and classes of entities and the possibility of manipulating them in terms of inherited properties. They are accessed with the standard query language OQL, which is the equivalent of SQL for relational databases, and are developed under standard language and representation (Unified Modeling Language, UML). They are well adapted for the storage of complex relation and the formulation of sophisticated queries and, in principle, enable more consistent developments of the databases, even if, for different technical reasons, they are not commonly used in molecular biology.

Inter-operativity of databases. Capacity of formulating the same operations in different databases with the same mechanisms (e.g. retrieving the corresponding EMBL entries during the inspection of a SWISS-PROT entry and formulating questions about the functions of the corresponding proteins and genes in the two databases).

Federated databases. A set of independent databases with proprietary data and different structures that agree on a protocol for exchanging information. The level of interchange can vary from simple data files to sophisticated answers and queries formulated on the fly.

Data warehouse. The overall structure for the creation of a repository of information. It does not necessarily imply a particular organisation of the information corresponding to the specific structure of the data.

Information retrieval. Accessing the documents, and other type of information repositories, that correspond to a given query. The Medline server at the NCBI is an example.

Information extraction. Selection of key information for a given question from repositories that have not been organised to answer that specific type of problem. Methods for selecting keywords from Medline abstracts are a first-level approach to information extraction in molecular biology, but other standard bioinformatics practices such as similarity searches in sequence databases can also be considered as information-extraction practices.

Linking/binding problem. The problem of maintaining (or finding!) the relations between comments in biological databases and the corresponding evidence. For example, finding the sentences in the abstracts from which an annotation about the function of a protein was derived.

Ontology. A defined set of entities (e.g. proteins) and their relations that tends to cover a full domain of knowledge (e.g. the localisation of all the proteins in cellular compartments). The most popular example of ontology in molecular biology is the GO ontology (Ashburner *et al.*, 2000) derived from deep biological knowledge, which unfortunately does not contain links to the evidence supporting the classification.

could be traced to its basic sentences in Medline abstracts, mainly because of the lack of coherence between protein and gene names used in databases and the literature. The lack of suitable pointers between annotations and facts has been criticised (Karp *et al.*, 2001), although efforts are under way to diminish this problem by including pointers to underlying information in key databases (Apweiler *et al.*, 2001).

The problem of binding annotations and sequences is ultimately due to the impossibility of accurately inferring function from sequence information. The common practice of comparing sequences and assigning the same function to similar novel proteins produces a considerable number of errors, since sequence similarity does not necessarily imply functional similarity. Furthermore, annotation errors accumulate due to the chain of assignments based on sequence similarities, which are particularly frequent because full genome sequences are now analysed automatically (Devos and Valencia, 2001). Subsequent re-annotation and elimination of errors is sometimes made unnecessarily difficult by legal considerations. Some biological databases are repositories in which the information is provided by the authors and can only be changed by them (e.g. EMBL; Stoesser *et al.*, 2001); other databases, such as SWISS-PROT (Bairoch and Apweiler, 2000), own the data they contain, and their database managers are responsible for managing the content.

The current panorama is even more

Instead of casting all the information in a single database, the most popular systems create links between existing ones

complex than the one of annotating the genome, since it requires linking and complementing the information about protein interactions and complexes, expression profiles, point mutations (SNPs) and metabolic profiles, which are distributed in different resources and repositories and require more sophisticated models than those developed for sequence and structure databases. For example, the enzymes, compounds and reactions in one metabolic pathway can be related to

other corresponding pathways and include compounds and reactions that occur in still other pathways. Numerous efforts are under way to construct

2000), in which the central repositories contain only the basic information but allow access to local repositories with additional data under previously agreed

There is an enormous gap between the information stored in >12 million papers published in the biomedical area and the comparatively small amount stored in even the most completely annotated databases

commonly accepted schemas and structures (e.g. array data; Brazma *et al.*, 2001) and metabolic pathways (<http://www.biopathways.org/>).

The most obvious solution is to integrate all this knowledge into a single database, upload all the information permanently and ensure its coherence—a daunting task indeed, which would need to deal with the continuous changes in independent databases, formal and semantic discrepancies between the various sources, and the lack of a generally accepted definition of concepts in molecular biology. This is clearly not feasible. Thus, instead of casting all the information in a single database, the most popular systems create links between existing ones. For example, SRS (Zdobnov *et al.*, 2002) creates extended text files containing all the information of each database and files with pointers relating entities in different databases; i.e. pointers between each SWISS-PROT entry and all corresponding entries containing the same gene in the EMBL database. This type of approach has obvious limitations and does not have the advantages offered by relational databases and managing systems described above. Moreover, they are difficult to install and maintain, as relations and pointers must be regularly updated. But it is simple and allows direct connection with the numerous databases on molecular biology that are distributed as flat text files and rapid access to data by following the links stored in these files.

A more sophisticated design to integrate information from different databases is the so-called 'federation of databases'. This concept does not require copying databases or duplicating data from one source to another but uses them to answer specific questions. Access to local databases and repositories is granted only in their own language and formats and only to the desired areas. This concept is similar to popular Napster servers (Butler,

conditions. In biology, this model is followed for the distributed annotation of the human genome (Hubbard *et al.*, 2002), a promising system the potential of which we have merely had the first glimpse. Along the same lines, the DiscoveryLink system (Haas *et al.*, 2001) includes facilities not only for accessing distributed repositories, but also for optimising queries in the native language of each database.

Clearly, the development of systems to store, analyse and allow access to information is a complex endeavour that is still lagging behind the speed of data

generation. But that is not all. Most relevant data are not actually stored in databases but in the scientific literature. There is an enormous gap between the information stored in >12 million papers published in the biomedical area—the critical knowledge that any biologist has accumulated—and the comparatively small amount stored in even the most completely annotated databases. And current efforts take only a fraction of this published information into account, if at all.

The first problem is access to bibliographic information, a widespread obstacle addressed by systems such as the popular web browser Google (<http://www.google.com>), which is able to retrieve documents, coded in different formats (e.g. HTML, XML, Postscript and PDF), that contain information relevant to a given query. The most obvious source of information in molecular biology is the National Library of Medicine's Medline repository of >12 million abstracts (<http://www.ncbi.nlm.nih.gov/pubmed>) that can be retrieved with various browser and information retrieval engines. Even so, the

Fig. 2. The future of information handling in biology. The progress in molecular biology, from the acquisition of information about single genes and proteins, the post-genomics technologies (DNA arrays, proteomics, etc.) and the integration of experiments about the complete system is confronted with the development in internet technologies essential for the manipulation of genomic information, from web-based systems, the new generation protocols for fast connections, to the GRID projects for integrating computing facilities using the internet connectivity. The developments in the internet and molecular biology are compared with the equivalent developments in field-specific databases, from the initial needs for database management systems (e.g. gene and protein sequences and protein structures) to the current efforts for the construction of databases dealing with complex proteomic and genomic data. The need for systems for the definition and classification of biological objects (ontology) are a key current development. Finally, implementation of knowledge management derived from the fields of information extraction and data mining are shown as key for the future of data management in molecular biology.

rest of the text is currently deposited in the private repositories of the editorial houses and is difficult to access for commercial and technical purposes. Several initiatives are under way to concentrate scientific texts in single repositories that can be explored and consulted simultaneously. In the US, the PubMed Central (<http://www.pubmedcentral.nih.gov/>) proposal addresses mainly the centralised deposition of publications and the issues related to property and legal rights. In Europe, the E-Oriel and E-BioSci (<http://www.e-bio-sci.org/>) initiatives (Grivell, 2002) focus on the technical aspects of accessing the publishers' servers in order to search for textual information with adequately unified queries and formats.

The second level of complexity is related to the extraction of the desired information from the retrieved documents. In the domain of molecular biology, the first information-extraction systems have appeared only recently and combine, to varying degrees, technologies related to statistics, information extraction, data mining and Natural Language Processing (for reviews, see Blaschke and Valencia, 2002, and Blaschke *et al.*, 2002). These applications seem to be able to accurately extract simple information such as protein names and interactions, even if it is difficult to compare them given the current lack of standardisation. This problem will also hamper any further progress, despite this apparent initial success. Moreover, text is only one part of the picture, in which other structures, such as web pages, figures and tables, play an increasingly important role and present difficulties far beyond those of the simple current systems.

It is clear that databases cannot account for all the knowledge in molecular biology. In the future, information-extraction and data-mining techniques will need to fill

the gaps by providing additional features directly retrieved from texts and other sources (Figure 2). This type of online system should be able to facilitate responses to the continually changing queries posed by the rapid evolution of biological research.

Acknowledgements

I would like to acknowledge the help of Jose Maria Fernandez-Gonzalez and Ramon Alonso-Allende during the preparation of the manuscript, as well as the interesting discussion and continuous technical update provided by them and other members of the Protein Design Group.

References

- Achard, F., Vaysseix, G. and Barillot, E. (2001) XML, bioinformatics and data integration. *Bioinformatics*, **17**, 115–125.
- Alonso-Allende, R., Fernandez-Gonzalez, J.M. and Valencia, A. (2002) The REGIA database (RegiaDB): status, limitations and future developments. *Comp. Func. Genom.*, in press.
- Andrade, M.A. *et al.* (1999) Automated genome sequence analysis and annotation. *Bioinformatics*, **15**, 391–412.
- Apweiler, R., Kersey, P., Junker, V. and Bairoch, A. (2001) Technical comment to "Database verification studies of SWISS-PROT and GenBank" by Karp *et al.* *Bioinformatics*, **17**, 533–534.
- Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
- Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- Blaschke, C. and Valencia, A. (2001) Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study. *Comp. Funct. Genom.*, **2**, 196–206.
- Blaschke, C. and Valencia, A. (2002) The peculiarities of the Molecular Biology nomenclature thwart the progress in information extraction. *IEEE Intell. Syst. Trends Controv.*, in press.
- Blaschke, C., Hirschman, L. and Valencia, A. (2002) Information extraction in Molecular Biology. *Brief. Bioinform.*, in press.
- Brazma, A. *et al.* (2001) Minimum information about a microarray experiment (MIAME): toward standards for microarray data. *Nature Genet.*, **29**, 365–371.
- Butler, D. (2000) Music software to come to genome aid? *Nature*, **404**, 694.
- Devos, D. and Valencia A. (2001) Intrinsic errors in genome annotation. *Trends Genet.*, **17**, 429–431.
- Grivell, L. (2002) Mining the bibliome: searching for a needle in a haystack? *EMBO rep.*, **3**, 200–203.
- Haas, L.M., Schwarz, P.M., Kodali, P., Kotlar, E., Rice, J.E. and Swope, W.C. (2001) DiscoveryLink: a system for integrated access to life sciences data sources. *IBM Syst. J.*, **40**, 489–511.
- Hubbard, T. *et al.* (2002). The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
- Karp, P.D., Paley, S. and Zhu, J. (2001) Database verification studies of SWISS-PROT and GenBank. *Bioinformatics*, **17**, 526–532; discussion 533–534.
- Stoesser, G. *et al.* (2001) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **29**, 17–21.
- Zdobnov, E.M., Lopez, R., Apweiler, R. and Etzold, T. (2002) The EBI SRS server: recent developments. *Bioinformatics*, **18**, 368–373.

Alfonso Valencia is at the Protein Design Group at the National Centre for Biotechnology CNB-CSIC in Madrid, Spain.
E-mail: valencia@cnb.uam.es

DOI: 10.1093/embo-reports/kvf104