

Comprehensive, comprehensible, distributed and intelligent databases: current status

Dmitrij Frishman¹, Klaus Heumann², Arthur Lesk³ and Hans-Werner Mewes¹

¹Munich Information Center for Protein Sequences of the German National Research Center for Environment and Health, Am Klopferspitz 18a, 82152 Martinsried, Germany, ²CuraGen Corporation, 555 Long Wharf Drive, New Haven, CT 06511, USA and ³Department of Haematology, University of Cambridge Clinical School, MRC Centre, Hills Road, Cambridge CB2 2QH, UK

Received on April 14, 1998; revised and accepted on May 6, 1998

Abstract

Motivation: *It is only a matter of time until a user will see not many but one integrated database of information for molecular biology. Is this true? Is it a good thing? Why will it happen? Where are we now? What developments are fostering and what developments are impeding progress towards this end?*

Supplementary information: *A list of WWW resources devoted to database issues in molecular biology is available at <http://www.mips.biochem.mpg.de>*

Contact: *frishman@mips.biochem.mpg.de*

Introduction

The problem of computer-assisted management of biological macromolecular sequence data is almost as old as the data themselves. In the early days, an article in *Science* by Dayhoff *et al.* (1980) described a nucleic acid data bank containing 200 entries and the total of 200 000 residues offered to the scientific community by telephone free of charge. Only a year later, *Nature* published a notice entitled 'Too many databanks?' (1981). Yet the latest special issue of *Nucleic Acids Research* lists 64 different databanks covering diverse areas of biological research, and nucleotide sequence data alone stand at over 1 billion bases (Benson *et al.*, 1998).

It is not only the flood and heterogeneity of biological data that make the issues of information representation, storage, structure, retrieval and interpretation critical and timely. There has also been a vast change in the user community. In the middle 1980s, fetching a PIR entry on a main-frame computer was an adventurous step that only few dared, and discovering an unexpected sequence similarity taken to be a rare bit of luck. Now, at the end of the 1990s, thousands of re-

searchers make use of biological databanks on a daily basis to answer routine questions, e.g. to find sequences similar to a newly sequenced gene, or to retrieve bibliographic references, or to investigate fundamental problems of modern biology (Koonin and Galperin, 1997). New technologies, of which the World Wide Web (WWW) has been the most revolutionary in terms of its impact on the daily practice of science (and other fields), have made it possible to create a high density of links between databanks. Database systems today are facing the task of serving ever increasing amounts of data of ever growing complexity to a user community that is growing nearly as fast as the data, and is getting more and more demanding.

The purpose of this review is to summarize the recent development of biological databases. We focus on technological aspects, such as models of data representation, data quality, available retrieval and data management software, interoperability and user interfaces. A number of related reviews exist, offering complementary points of view (Davidson *et al.*, 1995; Letovsky, 1995; Récipon and Malakowski, 1997; Baker and Brass, 1998). Detailed overviews of currently available Web resources for a molecular biologist can be found, for example, in Ashburner and Goodman (1997) and Galperin and Frishman (1998), and at many Web sites.

The PSB'98 database session

This article arises out of the database session of the Pacific Symposium for Biocomputing'98 (PSB'98); the theme of this session was our ability to identify and deliver high-quality information to the biological community. The contributions reflected developments in structuring and interlinking of databases (with pay-off in effective information retrieval); in automating the construction of entries (with, for example, pay-off in reduction in annotation errors); and the folding together of existing databases to produce alternative views of the corpus of data. Fujibuchi *et al.* (1998) gave a detailed description of their DBGET/LinkDB database retrieval sys-

A.L. and H.-W.M. are members of the CODATA Task Group on Biological Macromolecules.

tem for a wide range of molecular biology databases. Schulze-Kremer (1998) analyzed the possibility of using ontologies as a means to create consistent semantic relationships between data entries in different databases. Fukuda *et al.* (1998) developed an algorithm to extract biological material names (e.g. protein or gene names) from published articles and its potential applications for automatic sequence annotation. Wu and Shivakumar (1998) created a new protein family database by merging the superfamily information contained in the PIR-International database with the protein sequence motifs taken from the PROSITE database. M.Ponomarenko made a presentation of the ACTIVITY database (co-authors: J.V.Ponomarenko, A.S.Frolov and N.A.Kolchanov) that integrates various experimental and theoretical results on functional DNA and RNA sites.

Types of database content

The current scope of databases ranges from large-scale primary archiving projects down (or up, depending on your point of view) to individual, private, specialized 'boutique' collections serving the needs of particular user communities. These include the following.

General biological databanks: GenBank (Benson *et al.*, 1998), EMBL (Stoesser *et al.*, 1998), PIR-International (Barker *et al.*, 1998), SWISS-PROT (Bairoch and Apweiler, 1998), Online Mendelian Inheritance in Man (OMIM, 1997) and Protein Data Bank (Abola *et al.*, 1997).

Species-specific full-genome databases (not all completed yet) of the human and other organisms: *Saccharomyces cerevisiae* (Cherry *et al.*, 1998; Hodges *et al.*, 1998; Mewes *et al.*, 1998;), FlyBase (FlyBase Consortium, 1998) and a variety of small genomes (White and Kerlavage, 1996).

Databases specialized in subject matter, such as the database of transcription factors and their binding sites (TRANSFAC; Heinemeyer *et al.*, 1998) and the restriction enzyme resource (REBASE; Roberts and Macelis, 1998).

Derived databases containing added descriptive material on top of the primary data, or providing novel structuring of these data. Annotation is provided by automatic and/or manual methods. The most popular are the protein motif database (PROSITE; Bairoch *et al.*, 1997), structural classification of proteins (SCOP; Hubbard *et al.*, 1997), protein structure-sequence alignments (HSSP; Dodge *et al.*, 1998), protein domains (PFAM; Sonnhammer *et al.*, 1998) and conserved regions of proteins (BLOCKS; Henikoff *et al.*, 1998).

Special databases can grow and evolve very quickly as the result of the enormous data flow produced by automated sequencing and functional analysis. Whereas the large primary databases collect and collate 'atomic' information from the literature and from the scientific community, specialized data collections integrate, via curatorial expertise, information from a multiplicity of primary sources, including sequence,

structure, function, evolutionary relationships and bibliographic references. A rigid database classification has become obsolete, and the user chooses according to his individual needs from the rich WWW-accessible data palette.

Perhaps more significant than the growth in volume of the databases is the increasing complexity of the information available. Sequences are linked to structures, motifs and metabolic pathways. For example, databases of metabolic pathways (Karp, 1998) have intrinsically complex structures because the objects forming the data are nodes of networks linked by edges representing chemical reactions. The relationships between the nodes cannot be deduced from computable intrinsic properties like sequence homologies, but represent independent, non-sequence-related information. Another example of the structured nature of specialized databases is the database of Cytoskeletal Protein Interactions (Panzer *et al.*, 1997) which operates both with objects (e.g. proteins), their properties (e.g. protein classification) and possible relationships between objects in the form '\$a is a substrate of \$b' or '\$a regulates the binding of \$b to \$c'.

Intelligent data organization

The most commonly used methodology in computational molecular biology is comparison. Many biological objects come in families that share structural or functional features. Specifically for proteins, the concept of superfamilies was elaborated by Dayhoff (1976) and Zuckerkandl (1975) following the early observations that many proteins share sequence similarity, suggesting their common evolutionary origin. While in the early days of bioinformatics protein sequence and structure comparisons were made on a case-by-case basis, more recently many systematic efforts to organize protein-related information in the form of added-value data collections have been undertaken. The PIR-International data resource continued the Dayhoff tradition and maintains a database of similarity-based protein superfamily classifications (Barker *et al.*, 1996). Gonnet *et al.* (1992) made available the AIIAIDB database that contains all pairs of SWISS-PROT sequences with similarity above a certain threshold. For the PIR-International protein sequence database, a comprehensive, up-to-date FASTA homology database is available (Mewes *et al.*, 1997). A very successful attempt to cluster protein sequences is the Pfam database of protein domain families (Sonnhammer *et al.*, 1998) that contains manually curated and automatic components. Automatic approaches allow the delineation of the most conserved segments in related amino acid sequences that serve as characteristic signatures of protein families (Attwood *et al.*, 1998; Henikoff *et al.*, 1998). Through aligning protein sequences with sequences of known three-dimensional (3D) coordinates, it is possible to augment the amount of available structural information about proteins by an order of magni-

tude (Dodge *et al.*, 1998). Several classifications of protein 3D structures are available, produced both completely automatically (Orengo *et al.*, 1997; Holm and Sander, 1998) and by careful visual analysis (Hubbard *et al.*, 1997). Finally, Tatusov *et al.* (1997) provided a taxonomy of genes from multiple complete genomes. Beyond the organization of protein sequences following their evolutionary relationships, functional classifications, in particular for the completely sequenced genomes, have been compiled for *Escherichia coli* (Riley, 1993) and, in largely extended form, for the yeast genome (Mewes *et al.*, 1997).

Automating data annotation

It will hardly ever be possible completely to exclude human involvement from biological data annotation. However, mechanization of data acquisition dictates the necessity to look for ways to improve the productivity of the annotation teams. It is possible and desirable to automate reliably many annotation steps, or to automate the process partially by providing the biological experts with intelligently organized suggestive evidence.

The first large piece of molecular data to be subjected to computer-assisted annotation was the nucleotide sequence of the yeast chromosome III (Bork *et al.*, 1992). The set of programs used in this effort was the progenitor of the GeneQuiz genome analysis system (Scharf *et al.*, 1994; Casari *et al.*, 1996). Several other programs for large-scale sequence analysis have been developed (Gaasterland and Sensen, 1996; Frishman and Mewes, 1997; Walker and Koonin, 1997). Many research centers where primary data are being generated developed customized annotation systems tailored to their own technological processes (Eckman *et al.*, 1998). Typical features of such tools are systematic application of selected bioinformatics methods to sequence sets of any size, integration of all available evidence in the form of well-organized summaries for each data entry, application of hierarchical logical rules for producing functional and structural inferences with appropriate reliability estimates, and data storage, retrieval and visualization capabilities.

Methods are available that provide additional functional insights into biological sequence data without similarity comparisons. For example, des Jardins *et al.* (1997) described a way to delineate, with reasonable accuracy, enzyme EC numbers from easily computable protein sequence features through the application of machine intelligence approaches. Andrade and Valencia (1997) described a procedure to associate protein sequence data with bibliographic references stored in the MEDLINE database through frequency analysis of word occurrence.

No matter what program system is used, there are dangers inherent in automated annotation (Galperin and Koonin, 1998). Many molecular biology databanks are reluctant to

adopt automatic techniques, for fear of eroding annotation quality. One possible solution is to split a databank into two parts: the core section with carefully processed entries and a supplementary part for which the first-pass analysis is done automatically (Apweiler *et al.*, 1997).

Data quality and quality control by databanks

Databanks can, at two extremes, function as passive data repositories (archives), or as active reference compendia, issuing modifications of data and information content. Data in biological databanks contain facts, e.g. representations of biological macromolecules as strings or coordinates, and associated information which might be fuzzy, incomplete or subject to individual interpretation or conflicting nomenclature. Data quality has several elements: correctness, completeness, timeliness of capture, applied both to the newly measured properties, e.g. a new gene sequence, and the annotation.

Quality control should not be restricted to semantic checking of individual entries, but also include relationships to other parts of the database (George *et al.*, 1987). The growth in rates of data generation has implications for data quality. On the one hand, most new data entries are related to previously described objects and can inherit some part of their annotation. However, many newly determined data entries have no associated experimentally confirmed facts, and their annotation is based on predictions.

Two remarks about data quality are probably not in contention (perhaps the only two).

1. The quality of data in an archive can be no better than the data submitted to it, although processes of internal checking and exploitation of redundancy may help to identify and weed out data of poor quality.
2. Whatever else the databanks do, they should do no harm. We observe that databanks are tending to be more invasive in their approach to processing incoming data. Whereas databanks used to tend to function as repositories or archives, and act only passively in distributing the data to the community, they are now playing a more active role in interacting with the data. This interaction may involve checking procedures and/or addition of information in the form of annotations and links with other databanks. This is a concomitant of the development of structure in individual database entries and, of course, in the databases as a whole. In particular, genome-related databases actively curate and update data on a regular basis (e.g. MIPS and Stanford's SGD for the yeast genome).

The CODATA Task Group on Biological Macromolecules is undertaking a survey of quality control procedures in databanks in molecular biology. The types of data contained in macromolecular databanks differ in the extent to which er-

rors can be detected and corrected. For sequences, the general absence from the archive of 'raw data' (e.g. gels) makes it difficult to detect errors unless there are multiple determinations of the same sequence, or if detailed structural analysis of the protein corresponding to the sequence makes an error appear extremely likely (Bashford *et al.*, 1987), and the correct version of a sequence suspected of being erroneous is generally impossible to infer.

For protein and nucleic acid structures, in contrast, knowledge that the molecules must obey the general rules of stereochemistry and specialized rules of protein architecture makes it possible to try to evaluate the quality of a structure from the coordinate set alone. Several approaches to evaluating structures from coordinates alone have been proposed (Laskowski *et al.*, 1993; Hoofst *et al.*, 1997). In the case of both crystal and NMR structure determinations, moreover, implicit in the process of structure determination is the assembly of a set of experimental measurements as a data set in computer-readable form. For X-ray crystallography, this is the set or sets of structure factor magnitudes; for NMR, there is typically the set of assigned cross-peaks or even the spectral data themselves.

It is clear that effective error detection and correction of macromolecular structural data require the deposition and availability of these raw data, and it is very difficult to envisage any argument against requiring their deposition and distribution.

Given that databanks are taking a more active role in the assessment of data quality, what should be done with submissions that fail to meet desired standards? Databanks have always been under competing pressures to provide data quickly and completely, but also to aim for optimal data quality. One possibility is to withhold suspect data until they have been corrected (assuming that this is possible) and the other is to release them with a suitable warning. Personally, we tend to prefer the latter regime, but it is for the community as a whole to decide.

Data redundancy

Redundancy in primary molecular biology information arises as the consequence of parallel acquisition of the same or highly similar data from independent sources, inherent redundancy in the data itself, as well as small natural variation in the subjects or errors in the measurements that prevent the identification of essentially identical entries. Additional redundancy is often introduced by insufficiently coordinated annotation efforts. The situation is further complicated by the existence of two or more information resources in many subject areas, exemplified by two parallel protein sequence databanks: PIR-International (Barker *et al.*, 1998) and SWISS-PROT (Bairoch and Apweiler, 1998); and two parallel DNA

sequence databanks: GenBank (Benson *et al.*, 1998) and EMBL (Stoesser *et al.*, 1998).

Specifically for biopolymer sequence repositories, a number of mechanistic approaches have been elaborated to reduce the degree of data redundancy through exclusion of completely identical sequences [Bleasby and Wootton, 1990; the NRDB2 algorithm of Gish (unpublished)]. Sequence errors and variations cannot be taken into account. Combining all available protein sequence data from several sources in one non-redundant sequence collection leads to an ~2-fold reduction of the number of entries without losing any information and leads to substantial reduction in the computer resources necessary to perform similarity searches.

Perhaps the most severe case of data redundancy is represented by the EST (expressed sequence tag; Adams *et al.*, 1991) sequence collections. Multiple occurrence of nearly identical sequences in EST data collections is due to re-sampling of the same gene and is especially common for highly expressed genes. Several groups proposed methods to collapse clusters of related ESTs into distinct data entries representing individual genes, thus reducing this element of redundancy by several orders of magnitude (Adams *et al.*, 1995; Schuler *et al.*, 1996a; Hide *et al.*, 1997).

The need for a common language

Public databases distribute their contents as flat files, in some cases including indices for rapid data retrieval. In principle, all flat file formats are based on the organizational hierarchy of database, entry, record. Entries are the fundamental entities of molecular databases, but in contrast to the situation in the living cell that they purport to describe, database entries store objects in the form of atomic, isolated, non-hierarchical structures. Different databases may describe different aspects of the same biological unit, e.g. the nucleic acid and amino acid sequences of a gene, and the relationship between them must be established by links that are not intrinsically part of the data archives themselves.

The development of individual databases has generated a large variety of formats in their implementations. There is consensus that a common language, or at least that mutual intelligibility, would be a good thing, but this goal has proved difficult to achieve. Attempts to unify data formats have included application of Backus-Naur based syntax (George *et al.*, 1987), the development of an object-oriented database definition language (George *et al.*, 1993) and the use of Abstract Syntax Notation 1 (ASN.1; Ostell, 1990; Ohkawa *et al.*, 1995). None of these approaches has achieved the hoped for degree of acceptance.

Underlying the questions of mechanisms of intercommunication between databases of different structure and format is the need for common semantic standards and controlled vocabulary in annotations (see, for example, Pongor, 1988;

Rawlings, 1988). This problem is especially acute in comparative genomics. From the technological point of view, intergenome comparisons are interdatabase comparisons, which means that the databases to be compared have to speak the same language: keywords, information fields, weight factors, object catalogues, etc.

Perhaps the technical problems of standardization discussed in the preceding paragraphs could be addressed more easily in the context of a more general logical structure. As noted by Hafner and Fridman (1996), general biological data resources are data bases rather than knowledge bases: they describe miscellaneous objects according to the database schema, but no representation of general concepts and their relationships is given. Schulze-Kremer (1998) addressed this problem by developing ontologies for knowledge sharing in molecular biology. He proposed to create a repository of terms and concepts relevant to molecular biology, hierarchically organized by means of 'is a subset of' and 'is member of' operators.

Integrated data retrieval systems

An early attempt to provide a comprehensive interface to a variety of molecular biology databases was made by George and Orcutt in their ATLAS system (see George *et al.*, 1996). ATLAS generated data indices by parsing the heterogeneous formats into common fields allowing for cross-database multi-term queries avoiding the need for reformatting the database sources.

Over recent years, several highly sophisticated retrieval engines have emerged. Entrez (Schuler *et al.*, 1996b) is a gateway to the data collections maintained by NCBI. Those include nucleic acid and protein sequence data, 3D structures, genomes, taxonomic information, and the only freely available general literature database, PubMed. Perhaps the most powerful and unique feature of the Entrez system is that in addition to the static cross-references inherent to the underlying databases, extraction of related documents is also possible through the mechanism of 'neighbors'. For example, it is possible to retrieve bibliographic references related to the article you are currently looking at. This is done through lexicographical analysis of the abstract text and keywords. Sequence neighbors can be extracted by performing a BLAST similarity search. Entrez also has excellent graphical capabilities and is directly bundled with a protein structure viewer (Hogue *et al.*, 1996) and an extensive genome browser.

The Sequence Retrieval System (SRS; Etzold *et al.*, 1996), called 'the paragon of connectivity' (Brenner, 1995), allows the user to explore virtually all existing molecular biology databanks installed at different locations worldwide, making full use of the interdatabase links. The system includes a specially designed language for describing the structure of

databanks and the syntax of the data fields. The SRS user interface, which started as a system of pull-down menus on a VT100 terminal, has achieved in its Web incarnation an unprecedented degree of sophistication. Various Web query forms are adapted for interrogation by users' of different levels of expertise. Logical operators can be applied to field subsets; for instance, it is possible to ask for all sequences longer than 300 residues from *Escherichia coli*, containing the word kinase but not glucokinase in the description line, and published before 1995. Query forms can adapt themselves to the particular sets of databanks selected by the user. For example, the field for structure resolution will be offered if the PDB (or another structure databank) was selected. Output formats may also be customized, ranging from simple ID lists to complicated property views. Results of further computational analyses of retrieved sequence sets, including BLAST or CLUSTALW output, can also be indexed, linked, stored and accessed through the same interface.

Many cross-references provided in public databases are unidirectional and not mutual. In many cases, links between data collections are not direct, but involve multiple intermediate steps. SRS overcomes these difficulties by creating indices of all possible cross-links between all databases. This allows the use of all links bidirectionally, and permits the reaching of any databank from any other databank by the shortest path. If a new databank is added, it is sufficient to add a link to any of the existing nodes of the databank network.

Apart from the high technical level, several other features of SRS make it popular. First, it is relatively easy to install and maintain, which makes it suitable for running in-house bioinformatics Web sites. Second, it has a powerful command-line language allowing complex queries to be performed from programs and scripts. Alternatively, the C language API is available for creating calls to all SRS functions from C programs. Finally, anyone who creates a new databank can easily add it to the system by creating the appropriate description and indexing.

Another integrated retrieval system for molecular biology databanks, DBGET/LinkDB (Kanehisha, 1997) at Kyoto University and the Human Genome Center of Tokyo University, provides access to a variety of general data collections, as well as to the Kyoto Encyclopedia of Genes and Genomes.

A number of specialized retrieval systems for particular data types have been described. For example, Overton *et al.* (1994) developed an intelligent database query system (QGB) which is essentially a parser for the FEATURE TABLE field of the databases conforming to the DDJB/ENBL/GENBANK format. The system includes three components: the flat-file parser, the SQL-like query language, and the sequence entry parser itself, which analyses the logical structure of the FEATURE TABLE records and their relationships.

Alternatively, many WWW sites of interest provide useful facilities not based on a portable software solution, but offering carefully selected and organized data collections. For example, the popular Expasy WWW Server (Appel *et al.*, 1994) contains a wealth of information on virtually all aspects of computational molecular biology.

Database management systems

The retrieval systems described above are mainly user orientated, and access the databases in read-only mode. They are not suitable for database update and maintenance operations. Indices need to be rebuilt with each new data release and the interactive modification of data is not possible. These tasks are served by full-featured database management systems.

Relational database management systems (RDBMS) provide a sound and reliable model to implement databases in the form of tables representing records or record substructures, and are supported by extensive theory. The physical layer remains hidden from application programs and interactive users. RDBMS allow for fast database access through SQL (Structured Query Language) interfaces. Standard normalization rules avoid data redundancy and ensure referential integrity within large data sets. Database entries are built from tuples of tables, and specialized software tools are available to generate RDBMS tables from raw data. For example, programs exist for data acquisition from the Entrez system which parse entries in ASN.1 format and load them into Sybase SQL tables (Hart *et al.*, 1994; Korab-Laskowska *et al.*, 1998).

Commercial RDBMS are well-elaborated and widely distributed industrial applications that offer high data security and version control (e.g. Oracle, Sybase). One of the early applications of the RDBMS technology to biopolymer data was reported by Kanehisa *et al.* (1984). Now major database institutions (e.g. The European Bioinformatics Institute; Shomer *et al.*, 1996) employ RDBMS for their operation. The Genome Sequence Database (GSDB) at the National Center for Genome Resources (Harger *et al.*, 1998), powered by the Sybase software, even offers a direct SQL interface to 30 data tables describing various features of genetic sequences, such as annotated features, taxonomic information or bibliographic references. Other examples of RDBMS application for genetic data are the Mouse Genome Database (Blake *et al.*, 1998) and the *Bacillus subtilis* resource (Moszer *et al.*, 1995; Biaudet *et al.*, 1997).

One field where RDBMS has received much attention is the storage and representation of 3D macromolecular data (Islam and Sternberg, 1989; Huysmans *et al.*, 1991). Protein structures are suitable for representation in the form of tables, from the quaternary complex on the macro level, through subunit chains, super-secondary structures, secondary struc-

tures, down to individual residues, atom groups and atoms. The 3D base recently described by Abola *et al.* (1997) is an RDBMS for the protein structure databank also based on SYBASE. Higher level structures are built using the Object Protocol Model (Chen and Markowitz, 1995) to overcome RDBMS limitations for costly multiple table joins.

Disadvantages of RDBMS are the separation of the schema from the application software, which makes schema evolution difficult. In addition, hierarchical data structures are not directly supported, but in many cases are necessary to describe data in terms of relationships of groups of records, entry subsets, etc. In general, the relational approach has merely been successful for databanks which require schemas of limited complexity. Nevertheless, applications to protein structure were accomplished within a relational framework.

The alternative database design, based on object-oriented principles, emphasizes the tight coupling between data and the set of valid operations on that data. It allows dissection of a large database into components (such as references, sequences, annotation), with embodiment of methods into the data structures themselves, provides for flexible schema evolution, and permits the building of complex hierarchical data structures from classes by well-defined references to heterogeneous objects. Although commercial OODBMS have not yet matured and no widely available query language analogous to SQL exists, a number of applications to molecular biological data have been described, e.g. for a laboratory object-oriented information system (Goodman *et al.*, 1994) and for protein structure representation (Gray *et al.*, 1990). Ghosh (1998) reported an object-oriented database that can efficiently represent the complicated network of relationships between transcription factor polypeptides. An object-oriented database for sequence data processing and administration was developed at MIPS (A.Kaps, in preparation). AceDB (Durbin and Mieg, unpublished) is a special case: a general-purpose object-oriented database system developed specifically for genome projects. AceDB was clearly a milestone in making genome data accessible, especially for its extensive X-Windows-based graphics front-end.

Several biological data collections have implemented the recently emerged object-relational approach involving the application of a relational data storage engine for an object-oriented data model. Thus, the *Arabidopsis thaliana* Database at Stanford is powered by the Informix implementation of the object-relational DBMS (Flanders *et al.*, 1998). The basic concept appears to be attractive, since the reliable functions of the RDBMS can be employed to allow for object classes, hierarchies, etc. However, the necessary overhead for the assembly/disassembly process for objects and the separation of data and functions limits the use of object-relational hybrids.

Database interoperability and Internet technology

Biological data must be described in context rather than in isolation (Karp, 1996). Hence, many databases provide multiple links to other resources, but efficient use of these links requires intelligent retrieval systems. Attempts have been made to create interdatabase links automatically, restricted to few selected data resources, and with limited accuracy (Achard and Dessen, 1998). The user needs to be able to extract responses to a probe query from all possible sources, through a transparent and easy-to-use interface. The need for interoperability gave rise to the idea of an autonomous database federation (Robbins, 1994) through partial sharing of underlying database schema permitting cross-database queries, using, for example, SQL-based commercial database implementations. For example, attempts have been made to create a unified federated resource for microbiological information (Wertheim, 1995) suitable for intelligent interrogation. The prototype of such system is represented by the Ribosome Database Project (Maidak *et al.*, 1996).

An alternative approach is the concept of a warehouse, or a centralized data resource that manages a variety of data collections translated into a common format (Ritter, 1994). However, as those of us who live in Europe will readily understand, linking the community of databases through common semantics is impossible because of their extreme heterogeneity.

The recently emerged 'middleware' approach affords a chance to uncouple data access from data management and to allow for remote retrieval beyond the simple scripts fetching data from external databases. The most prominent industrial standard for a client-server based middleware is CORBA (Common Object Request Broker Architecture; Ben-Natan, 1995) as defined by the Object Management Group OMG. CORBA is a distributed object architecture that allows objects to communicate across networks through defined interfaces using the syntax of the Interface Definition Language (IDL). The object management architecture of CORBA specifies an application-level communication infrastructure.

Several CORBA-based applications have already appeared. Achard and Barillot (1997) suggest a set of interface definitions for molecular biology to access a simple but realistic data bank of Sequence Tag Sites. The European Commission supports a project to provide CORBA access to a set of public databases (EMBL, SWISS-PROT, PIR, TRANSFAC, and several others).

Stein *et al.* (1998) described an alternative approach to database interconnection. Their software system Jade establishes connection between the database servers and the application programs, and organizes data exchange through standardized relational tables and parameters. Information retrieved on the data server side is transformed into these

tables with the help of a specialized application called Jade adapter. Jade currently supports the AceDB, although incorporation of other database systems is anticipated.

The revolutionary rise of Internet technologies drives the present development of network programming tools. To overcome the limitations given by the available bandwidth and the server computing capacities, computing at the client's end is required. Client applications must be system independent since they are provided by Internet servers. The Internet programming language Java, developed by Sun Microsystems, is accepted as the common standard for programming slim applets that are transferred from server to client as bytecode and executed locally from the standard browser [see Stein (1996) for an introduction into applets for biologists]. Java allows efficient programming following the principle 'write once, run anywhere'. Current shortcomings of Java are slow execution times and virtual machine compatibility problems. One can expect that these obstacles will disappear since tools like the Java activator and Java just-in-time compilers will ensure that the applet runs in the client's browser environment with improved speed. This is backed by the Java Database Connectivity (JDBC) which allows for access to relational databases from Java programs or applets. The use of data resources in biology will benefit from the further development of the Internet software industry. We expect that biology will follow the general standardization trends. The next generation of Java tools will probably rely on Java Beans that are likely to be the key component of network services allowing independent applications to communicate. Java Beans will be compatible with standard communication protocols like CORBA or use Java remote method invocation (RMI) to connect distributed objects and enforce the development of compliant applications.

Visualization

Pictorial display, which has been recognized as essential for structural data for many years, is now recognized as also being necessary for large-scale genomic data. Visual tools are now important components of database systems and go far beyond merely replacing command-line queries through buttons and pull-down menus and displaying the retrieval results in a scrollable window. They allow the creation of complex views of large amounts of inter-related data, present various types of evidence in required context (e.g. genes together with regulatory elements), and increase the productivity of data mining (Robinson and Flores, 1997).

The development of visualization software was strongly motivated by genome sequencing projects. Genome data are especially rich in content, ranging from strain information, physical and genetic maps, clones, markers and chromosomes, to individual genes, protein catalogues and metabolic pathways. The Genome Division of NCBI offers an exten-

sive genome browser utilizing JavaScript to create expandable views of complete chromosomes and their fragments, with direct links to corresponding GenBank entries (Kuzio, 1996). More recently, a host of Java-based solutions to genome visualization emerged (Mewes *et al.*, 1997; Tamura *et al.*, 1997). The Genome Navigator by Grigoriev (1998) utilizes a Java applet called DerBrowser (Grigoriev, 1997) as a graphical gateway to a variety of remote data sources. Much attention has been paid to creating re-usable graphical software components that could be used in biological data visualization projects (Searls, 1995; the bioWidget Consortium: <http://goodman.jax.org/projects/biowidgets/>; Toldo, 1997). In particular, Java-based bioWidgets were utilized to create a powerful GenomeBrowser for the Berkeley Drosophila Genome Project. Junier and Bucher (1998) described a stand-alone Java applet for browsing protein and nucleic acid sequence data, intended, among other things, as a visual aid for sequence annotation.

Outlook

The molecular biology community faces an inundation of data. This implies that most of the contents of databanks will be recently determined data, and features, such as quality, will be characteristic of the newest methods of measurement. New experimental techniques will increase the amount and diversity of data; for example, the functional genomics, Proteome and expression profiles projects.

From the computational point of view, we are facing a non-static database. The data capture rates are going to be too fast to wait for human intervention to keep up with processing and annotation of entries in individual databanks, and with updating links between databases. At the same time, the need for more in-depth analysis in the context of the complete data set available will increase. As a consequence, the number of highly curated specialized databases will increase, adding entries to the future 1999 database issue of *Nucleic Acids Research*. As the number of these databases increases, interoperability between them becomes an even more critical issue. Secondary, dynamically curated databases must complement the heterogeneous public archives in a collaborative effort.

The technology will be driven by the standards of the consumer and multi-media industries. Novel software and communication tools will link data and programs, providing user-tailored views to the data and their interpretation. Beside the public databases, industrial efforts and commercial enterprises will try to fulfill the more specific needs of industrial applications.

The problem of the scientist will no longer be to get access to data, but to distill only those pieces of information that really contribute to the problem he or she is trying to solve. Staying on top of the development in one's field will become

harder. Databases should help their users accomplish this task, e.g. by alerting users to new data that matches their profile of interest. Pushing data to the scientist's desktop will more and more replace tedious net surfing.

We must develop and enlist software 'knowledge robots', called 'knowbots', that will cruise the Web, capturing and linking new data. It is possible to be optimistic that the knowbots will do the job fairly well (although probably not as well as humans), but a heavy responsibility lies on those who would create them.

Acknowledgements

We are grateful to Andrei Grigoriev for his invaluable comments on the manuscript. A.M.L. is supported by the Wellcome Trust.

References

- Abola, E.E., Sussman, J.L., Prilusky, J. and Manning, N.O. (1997) Protein Data Bank archives of three-dimensional macromolecular structures. *Methods Enzymol.*, **277**, 556–571.
- Achard, F. and Barillot, E. (1997) Ubiquitous distributed objects with CORBA. In *Pacific Symposium Biocomputing*, World Scientific, London, pp. 39–45.
- Achard, F. and Dessen, P. (1998) GenXref VI: automatic generation of links between two heterogeneous databases. *Bioinformatics*, **14**, 20–24.
- Adams, M.D. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.
- Adams, M.D. *et al.* (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, **377**, 3–174.
- Andrade, M.A. and Valencia, A. (1997) Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. *ISMB*, **5**, 25–32.
- Appel, R.D., Bairoch, A. and Hochstrasser, D.F. (1994) A new generation of information retrieval tools for biologists: the example of the exspasy www server. *Trends Biochem. Sci.*, **19**, 258–260.
- Apweiler, R. *et al.* (1997) Protein sequence annotation in the genome era: the annotation concept of SWISS-PROT+TREMBL. *ISMB*, **5**, 33–43.
- Ashburner, M. and Goodman, N. (1997) Informatics—genome and genetic databases. *Curr. Opin. Genet. Dev.*, **7**, 750–756.
- Attwood, T.K., Beck, M.E., Flower, D.R., Scordis, P. and Selley, J.N. (1998) The PRINTS protein fingerprint database in its fifth year. *Nucleic Acids Res.*, **26**, 304–308.
- Bairoch, A. and Apweiler, R. (1998) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Res.*, **26**, 38–42.
- Bairoch, A., Bucher, P. and Hofmann, K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res.*, **25**, 217–221.
- Baker, P.G. and Brass, A. (1998) Recent developments in biological sequence databases. *Curr. Opin. Biotechnol.*, **9**, 54–58.

- Barker, W.C. *et al.* (1998) The PIR-International protein sequence database. *Nucleic Acids Res.*, **26**, 27–32.
- Barker, W.C., Pfeiffer, F. and George, D.G. (1996) Superfamily classification in PIR-International protein sequence database. *Methods Enzymol.*, **266**, 59–71.
- Bashford, D., Chothia, C. and Lesk, A.M. (1987) Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.*, **196**, 199–216.
- Ben-Natan, R. (1995) *CORBA*. McGraw Hill, New York.
- Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J. and Ouellette, B.F. (1998) GenBank. *Nucleic Acids Res.*, **26**, 1–7.
- Biaudet, V., Samson, F. and Bessieres, P. (1997) Micado—a network-oriented database for microbial genomes. *Comput. Applic. Biosci.*, **13**, 431–438.
- Blake, J.A., Eppig, J.T., Richardson, J.E. and Davisson, M.T. (1998) The Mouse Genome Database (MGD): a community resource. Status and enhancements. The Mouse Genome Informatics Group. *Nucleic Acids Res.*, **26**, 130–137.
- Bleasby, A.J. and Wootton, J.C. (1990) Construction of validated non-redundant composite protein sequence databases. *Protein Eng.*, **3**, 153–159.
- Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. and Sonnhammer, E.L.L. (1992) Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III. *Protein Sci.*, **1**, 1677–1690.
- Brenner, S.E. (1995) World Wide Web and molecular biology. *Science*, **268**, 622–623.
- Casari, G., Ouzounis, C., Valencia, A. and Sander, C. (1996) GeneQuiz II: automatic function assignment for genome sequence analysis. In Hunter, L. and Klein, T.E. (eds), *First Annual Pacific Symposium on Biocomputing*. World Scientific, Hawaii, pp. 707–709.
- Chen, I.A. and Markowitz, V.M. (1995) An overview of the Object-Protocol Model (OPM) and OPM data management tools. *Inf. Syst.*, **20**, 393–418.
- Cherry, J.M. *et al.* (1998) SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.*, **26**, 73–79.
- Davidson, S.B., Overton, C. and Buneman, P. (1995) Challenges in integrating biological data sources. *J. Comput. Biol.*, **2**, 557–572.
- Dayhoff, M.O. (1976) The origin and evolution of protein superfamilies. *Fed. Proc.*, **35**, 2132–2138.
- Dayhoff, M.O., Schwartz, R.M., Chen, H.R., Hunt, L.T., Barker, W.C. and Orcutt, B.C. (1980) Nucleic acid sequence bank. *Science*, **209**, 1182.
- des Jardins, M., Karp, P.D., Krummenacker, M., Lee, T.J. and Ouzounis, C.A. (1997) Prediction of enzyme classification from protein sequence without the use of sequence similarity. *ISMB*, **5**, 92–99.
- Dodge, C., Schneider, R. and Sander, C. (1998) The HSSP database of protein structure–sequence alignments and family profiles. *Nucleic Acids Res.*, **26**, 313–315.
- Eckman, B.A., Aaronson, J.S., Borkowski, J.A., Bailey, W.J., Elliston, K.O., Williamson, A.R. and Blevins, R.A. (1998) The Merck Gene Index browser: an extensible data integration system for gene finding, gene characterization and EST data mining. *Bioinformatics*, **14**, 2–13.
- Etzold, T., Ulyanov, A. and Argos, P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.
- Flanders, D.J., Weng, S., Petel, F.X. and Cherry, J.M. (1998) AtDB, the *Arabidopsis thaliana* database and graphical-web-display of progress by the *Arabidopsis* Genome Initiative. *Nucleic Acids Res.*, **26**, 80–84.
- FlyBase Consortium (1998) FlyBase: a *Drosophila* database. *Nucleic Acids Res.*, **26**, 85–88.
- Frishman, D. and Mewes, H.W. (1997) PEDANTic genome analysis. *Trends Genet.*, **13**, 415–416.
- Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y. and Kanehisa, M. (1998) DBGET/LinkDB: an integrated database retrieval system. In Altman, R.B., Dunker, A.K., Hunter, L. and Klein, T.E. (eds), *Pacific Symposium on BIOCOMPUTING'98*. World Scientific, London, pp. 683–694.
- Fukuda, K., Tamura, A., Tsunoda, T. and Takagi, T. (1998) In Altman, R.B., Dunker, A.K., Hunter, L. and Klein, T.E. (eds), *Pacific Symposium on BIOCOMPUTING'98*. World Scientific, London, pp. 707–718.
- Gaasterland, T. and Sensen, C.W. (1996) MAGPIE: automated genome interpretation. *Trends Genet.*, **12**, 76–78.
- Galperin, M.Y. and Frishman, D. (1998) Toward automated prediction of protein function from microbial genomic sequences. *Methods Microbiol.*, in press.
- Galperin, M.Y. and Koonin, E.V. (1998) Sources of systematic errors in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol.*, **1**, 0007.
- George, D.G., Mewes, H.-W. and Kihara, H. (1987) A standardized format for sequence data exchange. *Protein Seq. Data Anal.*, **1**, 27–39.
- George, D.G., Orcutt, B.C., Mewes, H.-W. and Tsugita, A. (1993) An object-oriented sequence database definition language (sddl). *Protein Seq. Data Anal.*, **5**, 357–399.
- George, D.G., Barker, W.C., Mewes, H.-W., Pfeiffer, F. and Tsugita, A. (1996) The PIR-International protein sequence database. *Nucleic Acids Res.*, **24**, 17–20.
- Ghosh, D. (1998) OOTFD (Object-Oriented Transcription Factors Database): an object-oriented successor to TFD. *Nucleic Acids Res.*, **26**, 360–362.
- Gonnet, G.H., Cohen, M.A. and Benner, S.A. (1992) Exhaustive matching of the entire protein sequence database. *Science*, **256**, 1433–1445.
- Goodman, N., Rozen, S. and Stein, L. (1994) In *Proceedings of the 20th International Conference on Very Large Data Bases*, Santiago, Chile, pp. 722–729.
- Gray, P.M.D., Paton, N.W., Kemp, G.J.L. and Fothergill, J.E. (1990) An object-oriented database for protein structure analysis. *Protein Eng.*, **3**, 235–243.
- Grigoriou, A. (1997) Genomes with a view. *Trends Genet.*, **13**, 499.
- Grigoriou, A. (1998) Genome Navigator. *Trends Microbiol.*, **6**, 184.
- Hafner, C.D. and Fridman, N. (1996) Ontological foundations for biology knowledge models. *ISMB*, **4**, 78–87.
- Harger, C. *et al.* (1998) The Genome Sequence DataBase (GSDB): improving data quality and data access. *Nucleic Acids Res.*, **26**, 21–26.
- Hart, K.W., Searls, D.B. and Overton, G.C. (1994) SORTEZ: a relational translator for NCBI's ASN.1 database. *Comput. Applic. Biosci.*, **10**, 369–378.

- Heinemeyer, T. *et al.* (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.*, **26**, 362–367.
- Henikoff, S., Pietrovski, S. and Henikoff, J.G. (1998) Superior performance in protein homology detection with the Blocks Database servers. *Nucleic Acids Res.*, **26**, 309–312.
- Hide, W., Burke, J., Christoffels, A. and Miller, R. (1997) Toward automated prediction of protein function from microbial genomic sequences. In Miyano, S. and Takagi, T. (eds), *Genome Informatics*. Universal Academy Press, Tokyo, pp. 187–196.
- Hodges, P.E., Payne, W.E. and Garrels, J.I. (1998) The Yeast Protein Database (YPD): a curated proteome database for *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **26**, 68–72.
- Hogue, C.W., Ohkawa, H. and Bryant, S.H. (1996) A dynamic look at structures: WWW-Entrez and the Molecular Modeling Database. *Trends Biochem. Sci.*, **21**, 226–229.
- Holm, L. and Sander, C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.*, **26**, 316–319.
- Hooft, R.W., Sander, C. and Vriend, G. (1997) Objectively judging the quality of a protein structure from a Ramachandran plot. *Comput. Applic. Biosci.*, **13**, 425–430.
- Hubbard, T.J.P., Murzin, A.G., Brenner, S.E. and Chothia, C. (1997) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **25**, 236–239.
- Huysmans, M., Richelle, J. and Wodak, S.J. (1991) SESAM: a relational database for structure and sequence of macromolecules. *Proteins*, **11**, 59–76.
- Islam, S.A. and Sternberg, M.J.E. (1989) A relational database of protein structures designed for flexible enquiries about conformation. *Protein Eng.*, **2**, 431–442.
- Junier, T. and Bucher, P. (1998) SEView: a Java applet for browsing molecular sequence data. *In Silico Biol.*, **1**, 0003.
- Kanehisa, M. (1997) A database for post-genome analysis. *Trends Genet.*, **13**, 375–376.
- Kanehisa, M., Fickett, J.W. and Goad, W.B. (1984) A relational database system for the maintenance and verification of the Los Alamos sequence library. *Nucleic Acids Res.*, **12**, 149–158.
- Karp, P.D. (1996) Database links are a foundation for interoperability. *Trends Biotechnol.*, **14**, 274–279.
- Karp, P.D. (1998) Metabolic databases. *Trends Biochem.*, **23**, 114–116.
- Koonin, E.V. and Galperin, M.Y. (1997) Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Curr. Opin. Genet. Dev.*, **7**, 757–763.
- Korab-Laskowska, M., Rioux, P., Brossard, N., Littlejohn, T.G., Gray, M.W., Lang, B.F. and Burger, G. (1998) The Organelle Genome Database Project (GOBASE). *Nucleic Acids Res.*, **26**, 138–144.
- Kuzio, J. (1996) Genomic sequence presentation. *Trends Genet.*, **12**, 321–322.
- Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.*, **26**, 283–291.
- Letovsky, S. (1995) Beyond the information maze. *J. Comput. Biol.*, **2**, 539–546.
- Maidak, B.L., Olsen, G.J., Larsen, N., Overbeek, R., McCaughey, M.J. and Woese, C.R. (1996) The ribosomal database project (RDP). *Nucleic Acids Res.*, **24**, 82–85.
- Mewes, H.W., Albermann, K., Heumann, K., Liebl, S. and Pfeiffer, F. (1997a) MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res.*, **25**, 28–30.
- Mewes, H.W. *et al.* (1997b) The yeast genome directory. *Nature*, **387**, 7–65.
- Mewes, H.W., Hani, J., Pfeiffer, F. and Frishman, D. (1998) MIPS: a database for protein sequences and complete genomes. *Nucleic Acids Res.*, **26**, 33–37.
- Moszer, I., Glaser, P. and Danchin, A. (1995) Subtilist: a relational data base for the bacillus subtilis genome. *Microbiology*, **141**, 261–268.
- Ohkawa, H., Ostell, J. and Bryant, S. (1995) MMDB: an ASN.1 specification for macromolecular structure. *ISMB*, **3**, 259–267.
- OMIM (1997) *Online Mendelian Inheritance in Man, OMIM (TM)*. Center for Medical Genetics, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). Available at <http://www.ncbi.nlm.nih.gov/omim/>
- Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) CATH—A hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
- Ostell, J. (1990) *GenInfo ASN.1 Syntax: Sequences*. NCBI Technical Report Series, National Library of Medicine, NIH. Technical Report Number 1, p. 37.
- Overton, G.C., Aaronson, J.S., Haas, J. and Adams, J. (1994) Qgb: a system for querying sequence database fields and features. *J. Comput. Biol.*, **1**, 3–14.
- Panzer, S., Cooley, L. and Miller, P.L. (1997) Using explicitly represented biological relationships for database navigation and searching via the World-Wide Web. *Comput. Applic. Biosci.*, **13**, 281–290.
- Pongor, S. (1988) Novel databases for molecular biology. *Nature*, **332**, 24–24.
- Rawlings, C.J. (1988) Designing databases for molecular biology. *Nature*, **334**, 477–477.
- Réçipon, H. and Makalowski, W. (1997) The biologist and the World Wide Web: an overview of the search engines technology, current status and future perspectives. *Curr. Opin. Biotechnol.*, **8**, 115–118.
- Riley, M. (1993) Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.*, **57**, 862–952.
- Ritter, O. (1994) The integrated genomic database. In Suhai, S. (ed.), *Computational Methods in Genome Research*. Plenum, New York, pp. 57–73.
- Robbins, R.J. (1994) Genome informatics I: community databases. *J. Comput. Biol.*, **1**, 173–190.
- Roberts, R.J. and Macelis, D. (1998) REBASE—restriction enzymes and methylases. *Nucleic Acids Res.*, **26**, 338–350.
- Robinson, A.J. and Flores, T.P. (1997) Novel techniques for visualizing biological information. *ISMB*, **5**, 241–249.
- Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C. and Sander, C. (1994) GeneQuiz: a workbench for sequence analysis. *ISMB*, **2**, 348–353.
- Schuler, G.D. *et al.* (1996a) A gene map of the human genome. *Science*, **274**, 540–546.
- Schuler, G.D., Epstein, J.A., Ohkawa, H. and Kans, J.A. (1996b) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.

- Schulze-Kremer,S. (1998) Ontologies for molecular biology. In Altman,R.B., Dunker,A.K., Hunter,L. and Klein,T.E. (eds), *Pacific Symposium on BIOCOMPUTING'98*. World Scientific, London, pp. 695–706.
- Searls,D.B. (1995) bioTk:componentry for genome informatics graphical user interfaces. *Gene*, **163**, GC1–16.
- Shomer,B., Harper,R.A. and Cameron,G.N. (1996) Information services of the European Bioinformatics Institute. *Methods Enzymol.*, **266**, 3–27.
- Sonnhammer,E.L.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, **26**, 322–325.
- Stein,L. (1996) Web applets: Java, JavaScript and ActiveX. *Trends Genet.*, **12**, 484–485.
- Stein,L.D., Cartinhour,S., Thierry-Mieg,D. and Thierry-Mieg,J. (1998) JADE: An approach for interconnecting bioinformatics databases. *Gene*, **209**, 39–43.
- Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor,M. and Sterk,P. (1998) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **26**, 8–15.
- Tamura,T., Mori,H. and Sugawara,H. (1997) Genome Information Broker for large and small genomes. *Trends Genet.*, **13**, 498.
- Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
- Toldo,L.I. (1997) JaMBW 1.1: Java-based Molecular Biologists' Workbench. *Comput. Applic. Biosci.*, **13**, 475–476.
- Walker,D.R. and Koonin,E.V. (1997) SEALS: a system for easy analysis of lots of sequences. *ISMB*, **5**, 333–339.
- Wertheim,M. (1995) Call to desegregate microbial databases. *Science*, **269**, 1516.
- White,O. and Kerlavage,A.R. (1996) TDB: new databases for biological discovery. *Methods Enzymol.*, **266**, 27–40.
- Wu,C.H. and Shivakumar,S. (1998) Proclass protein family database: new version with motif alignments. In Altman,R.B., Dunker,A.K., Hunter,L. and Klein,T.E. (eds), *Pacific Symposium on BIOCOMPUTING'98*. World Scientific, London, pp. 719–730.
- Zuckerklund,E. (1975) The appearance of new structures and functions in proteins during evolution. *J. Mol. Evol.*, **7**, 1–57.