Macromolecular structure information and databases

The EU BRIDGE DataBase Project Consortium*

The current status and future outlook of macromolecular structure databases and information handling, with particular reference to European databases, are reviewed. Issues concerning the efficiency with which data are represented, validated, archived and accessed are discussed in view of the fast growing body of information on structures of biological macromolecules.

PROGRESS IN GENETIC engineering, X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy, and the advent of cheap and powerful computers, have brought about an exponential growth of data on macromolecular three-dimensional structures. Although the number of protein structures known today (approximately 3500) is still well below that of protein sequences (over 50 000), its growth rate parallels that observed for protein sequences several years ago and continues to rise, with an expected 30 000 known structures by the turn of this century. The expected number of nucleic acid structures is smaller, but will also rise steadily.

Managing the information on biomacromolecule sequences and structures has become one of the major challenges of modern molecular biology. This is particularly difficult in the case of protein three-dimensional structures, because it is necessary to handle a wealth of complex information, some of which must be derived (or calculated) from the atomic coordinates by applying specialized programs (Box 1). Furthermore, the data must be validated, i.e. the quality and the consistency of the data must be assessed. To promote scientific progress, structural information must also be integrated with other data, such as that on function, gene structure and phylogeny, which are obtained from other sources or stored in different databanks. In addition, the data must be easily and widely accessible to the molecular biology community.

Meeting this challenge requires that the ways of storing, cross-referencing and accessing information must be enhanced, which in turn requires that the data are organized in ways that can describe items such as atoms, residues, secondary structure elements, folding motifs, domains, subunits and crystal contacts, as well as the complex relationships between them.

Here, we give an overview of the recent advances in the area of macromolecular structure databases, with illustrations taken from works by the European BRIDGE DataBase Project Consortium.

*Peter M. D. Gray and Graham J. L. Kemp are at the Computing Science Department, University of Aberdeen, Aberdeen, UK AB9 2UE.
Christopher J. Rawlings and Nigel P. Brown are at the ICRF, 44 Lincoln's Inn Fields, London, UK WC2A 3PX.
Christian Sander is at the EMBL, 69012 Heidelberg, Meyerhofstraße 1, Germany.
Janet M. Thornton and Christine M. Orengo are at the Biochemistry and Molecular Biology Department, University College London, London, UK WC1E 6BT.
Shoshana J. Wodak and Jean Richelle are at the UCMB, Université Libre de Bruxelles, av. F. D. Roosevelt 50 – CP160/16, 1050 Bruxelles, Belgium.

Current status of macromolecular structure resources

Publicly available resources for storing, retrieving and manipulating information on macromolecular structures can be divided into the archives for the primary data (obtained from the authors or in the literature) and the 'derived data' resources (Box 1). Major resources are listed in Table I [a World Wide Web (WWW) version of this table can be found at http://www.ucmb.ulb.ac.be/StructResources].

The primary data are archived at the Protein Structure Databank (PDB) at Brookhaven National Laboratories (BNL)[1] and the Nucleic Acid Database (NDB)[2] and are freely available to all. Many countries and laboratories also keep a mirrored copy of the PDB to improve local access.

The 'derived data' resources can be grouped according to the mode used to access them: some data are stored in ASCII files, available by file transfer; others can be 'browsed' on the WWW using tools such as Mosaic™ or Netscape™, which allow limited navigation through 'pre-canned' data or reports; the most sophisticated access is available when the data are organized in a database (under management systems such as ORACLE® and SYBASE®). For data retrieval, databases have the advantage of allowing for custom-designed queries to be made without writing specialized programs.

An important consideration for efficient access to information is the availability of links between different data resources. Currently in PDB files, no

Box 1. Derived data

Computed data from a single three-dimensional structure

- Backbone and side chain dihedral angles.
- Atom connectivities.
- H-bonding partners.
- Secondary structure assignments.
- Topological description of β-sheets.
- Topological and geometric description of the relative arrangement of all secondary structures.
- Location of specific structure motifs: specific types of turns, supersecondary structures.
- Solvent accessible surface areas of atoms, residues, whole subunits and multi-subunit complexes.
- Volumes of atoms, residues, whole subunits and multi-subunit complexes.
- Locations of empty and water filled cavities; their volume, surface area.
- Residue–residue contact maps.
- Limits of structural domains.
- List of residues involved in subunit contacts.
- List of residues involved in ligand binding.
- List of residues involved in crystal contacts.
- Pointers to homologous structures.
- Pointers to homologous sequences in other databanks.
- Annotation of structure quality assessment.

Data computed from surveys of many structures

- Preferred side chain rotamers.
- Preferred modes of protein sidechain interactions.
- Preferred modes of protein–water interactions.
- Repertoire of turn motifs and their characteristics.
- Repertoire of sheet topologies.
- Repertoire of helix–helix, helix–sheet and sheet–sheet interactions.
- Repertoire of folding motifs.

PII: S0968-0004(96)10037-2

Table I. Macromolecular structure information resources

Resources of primary data		
PDB	Protein Data Bank at Brookhaven National Laboratories[1]	http://www.pdb.bnl.gov
NDB	Nucleic Acid Database[2] – Rutgers University (USA) – European mirror at the EBI (UK)	 http://ndbserver.rutgers.edu http://www.ebi.ac.uk/NDB
BioMagResBank	Protein, peptide and nucleic acid NMR spectroscopy database[17]	http://www.bmrb.wisc.edu
Useful collections of data		
HSSP	Homology-derived Secondary Structure of Proteins[18]	ftp://ftp.embl-heidelberg.de/pub/databases/protein_extras/hssp
3D_ALI	Sequence alignments of structurally superposed proteins[19]	http://www.embl-heidelberg.de/argos/ali/ali.html
PUU	Putative protein structural domains[20]	ftp://ftp.embl-heidelberg.de/pub/databases/protein_extras/puu/domains.puu
Walsh annot.	Annotation of PDB entries[21]	http://www.scs.uiuc.edu/~lwalsh/files.info.llw
Sites to browse and search on the World Wide Web		
PDBBrowse	PDB Browser to access the structure database	http://www.pdb.bnl.gov/cgi-bin/browse
NDB	Query Interface of the Nucleic Acid Database – at Rutgers University (USA) – at EBI mirror (UK)	 http://ndbserver.rutgers.edu/interface/index.html http://www.ebi.ac.uk/NDB/interface/index.html
MOOSE	Macromolecular Structure Query of the PDB ftp archive	http://www.sdsc.edu/moose
Molecules R US	Forms interface to search an index of the PDB	http://molbio.info.nih.gov/cgi-bin/pdb
PBDSummaries	Summary information and images on each PDB entry	http://www.biochem.ucl.ac.uk/bsm/pdbsum/index.html
SCOP	Structural Classification of Protein[22]	http://scop.mrc-lmb.cam.ac.uk/scop
CATH	Protein Structure Classification	http://www.biochem.ucl.ac.uk/bsm/cath/CATHintro.html
FSSP	Fold classification based on Structure–Structure alignment of Proteins[23]	http://www.embl-heidelberg.de/dali/fssp/fssp.html
LPFC	A Library of Protein Family Cores	http://camis.stanford.edu/projects/helix/LPFC
SRSWWW	A Browser for Databanks in Molecular Biology[24]	http://www.ch.embnet.org/srs/index.html
SWISS-3DIMAGE	High quality pictures of biological macromolecules[25]	http://expasy.hcuge.ch/sw3d/sw3d-top.html
Entrez	Access now MMDB, a database of protein structures	http://www3.ncbi.nlm.nih.gov/Entrez
Protein Motions	A database of domain, loop and subunit motions[26]	http://hyper.stanford.edu/~mbg/ProtMotDB
Structure databases		
IDITIS	Relational database under proprietary RDBMS[27]	http://www.oxmol.co.uk/PRODUCTS/iditis_top.html
NDB	Relational database under Sybase[2]	http://ndbserver.rutgers.edu
OOPDB	Object-oriented database under ObjectStore[28]	mail to bourne@sdsc.edu
WPDB	Compressed and indexed PDB under MS Windows[29]	http://www.sdsc.edu/CCMS/Packages/wpdb.html
P/FDM	Object-oriented database[5]	http://www.csd.abdn.ac.uk/~pfdm
PKB	Object-oriented database under S[10]	ftp://ncbi.nlm.nih.gov/pub/pkb
ProLink	Relational database under Sybase[30]	http://bmerc-www.bu.edu/plforms/forms-toc.html
PROTEP	PROLOG database	mail to chris_rawlings-1@sbphrd.com
SESAM	Relational database under Sybase[12]	mail to shosh@ucmb.ulb.ac.be

explicit references exist to data entries in other resources, although we expect this will change in the future. However, many other resources have included explicit links to the PDB files. For example, in the SWISS-PROT sequence databank, there is a linker in each sequence entry, which gives the accession number for the corresponding PDB entry, if available. Some of the important primary archives, which incorporate such explicit links to the structural data, are also listed in Table I.

Standards for data representation

To process efficiently the rapidly increasing body of data on the structures of biological macromolecules there is a need for accepted standards for unambiguously describing all the relevant information on macromolecular structures, which can form the basis for the exchange and storage of data as computer files. Such standards will need to be richer than the currently used PDB file format, which has been the *de facto* standard for the last 20 years. Ideally, they should have the capacity to encode not only the atomic coordinates and chemical structure, but also information on the experimental procedure and data (X-ray diffraction or NMR), information on biological function, structural descriptions (e.g. location of domains), links to other databases and any relevant derived data, including criteria for evaluating the accuracy of the model. While all these data might not necessarily be available for all structures, an agreed basic and minimal set of data items should be defined that allows internally consistent structure descriptions to be produced. All this would have to be encoded in a way that can easily be read by computer programs, and that allows for automatic consistency and error checking. It must also facilitate the exchange of data and software, allow views of the data, which can be easily mapped into a data structure held in computer memory or a database, and

Table I. Macromolecular structure information resources (contd)

Other databanks/databases		
BMCD	The Biological Macromolecule Crystallization Database[32]	http://ibm4.carb.nist.gov:4400/bmcd/bmcd.html
Klotho	Biochemical Compounds Declarative Database (stereochemical configurations)	http://ibc.wustl.edu/klotho
Cross-links with other databanks/databases		
SWISS-PROT	Annotated protein sequence databank[33]	http://expasy.hcuge.ch/sprot/sprot-top.html
PIR	Protein Identification Resource – protein sequence databank[34]	http://www.gdb.org/Dan/proteins/pir.html
ATLAS	Retrieval program to access sequences databases at MIPS	http://www.mips.biochem.mpg.de/mips/programs/atlas.html
PROT-FAM	Browser to see the protein classification tree at MIPS	http://www.mips.biochem.mpg.de/mips/programs/classification.html
GenBank	The genetic sequence databank[35]	http://www3.ncbi.nlm.nih.gov/Entrez
ENZYME	The enzyme nomenclature databank[36]	http://expasy.hcuge.ch/sprot/enzyme.html ftp://expasy.hcuge.ch/databases/enzyme
Enzyme Structures Database	Another enzyme database	http://www.biochem.ucl.ac.uk/bsm/enzymes/index.html
NRL-3D	A sequence database derived from the PDB[37]	http://www.gdb.org/Dan/proteins/nrl3d.html
ProDom	A database of protein or domain families[38]	http://www.sanger.ac.uk/~esr/prodom.html http://protein.toulouse.inra.fr/prodom.html
PROSITE	A databank of biological sites, patterns and profiles[39]	http://expasy.hcuge.ch/sprot/prosite.html
PRINTS	Protein Motif Fingerprint Database[40]	http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html
BLOCKS	A database of highly conserved regions in protein sets[41]	http://www.blocks.fhrcr.org
MOTIFS	Searching Protein Sequence Motifs	http://www.genome.ad.jp/SIT/MOTIF.html
Compilations of resources on the WWW		
Structural Biology and Engineering		http://www.cryst.bbk.ac.uk/CEC/biomol.html
ExPASy	Expert Protein Analysis System	http://expasy.hcuge.ch
DbBrowser	Access to a composite sequence database and its derived database of fingerprints[42]	http://www.biochem.ucl.ac.uk/bsm/dbbrowser
EBI	European Bioinformatics Institute server	http://www.ebi.ac.uk
Bio-wURLd	A searchable collection of URLs related to bioinformatics, biochemistry and molecular biology	http://www.ebi.ac.uk/htbin/bwurld.pl
Pedro's home page	BioMolecular Research Tools	http://www.public.iastate.edu/~pedro/research_tools.html http://www.biophys.uni-duesseldorf.de/bionet/research_tools.html http://www.fmi.ch/biology/research_tools.html http://www.peri.co.jp/Pedro/research_tools.html
Structural Information Resources	The compilation of resources listed in this table	http://www.ucmb.ulb.ac.be/StructResources

leave room for future expansion (new data fields and other application areas).

At the National Center for Biotechnology Information (NCBI), the problem of exchange of data and information between the sequence, structure and citation databases has been addressed by using ASN.1 (Abstract Syntax Notation One)[3], a general purpose ISO standard interchange format previously developed to transmit simple, as well as structured, data over the electronic network independently of their representation in the computers.

The International Union of Crystallography (IUCr) has adopted the more specialized CIF (Crystallographic Information File) standard to represent structural data for small molecules. A macromolecular extension to this standard, mmCIF, has recently been developed by an international working group, which includes protein crystallographers, NMR spectroscopists, protein modelers and representatives of several international bodies. The extension mmCIF embodies many of the requirements described above (P. E. Bourne *et al.*, unpublished) and is therefore a major breakthrough for the macromolecular structure field; the primary WWW site for mmCIF is at http://ndbserver.rutgers.edu/mmcif; this contains the dictionary in text and hypertext mark-up language (HTML), an application program interface to CIF (SIFLIB), a growing number of software tools and an mmCIF tutorial from http://www.sdsc.edu/CompSci/pb/cif/ci.html. The widespread adoption of the mmCIF and ASN.1 standards will depend crucially on the availability of public domain software to process and validate the information.

Validation of structural data

A database is useful only if the data it contains are accurate. Therefore, it is most important that the atomic coordinates and the experimental data deposited in protein structure databases be internally consistent and properly validated. Macromolecular coordinates are, in essence, only models derived to give the best fit to the experimental data, be they electron densities or NMR-derived distance constraints. The experimental data, however, might not be sufficiently good to define the model accurately, and furthermore, the accuracy might not be the same everywhere in the structure. Some measure of the

Table II. Quality measures

	Global	Local: by residue or atom
X-ray		
Quality of data	Resolution	
Fit of data to model	R-factor (free R-factor)[a]	Occupancy[b]; B-value[c]
NMR		
Quality of data	Number of constraints/residues[d]	
Fit of data to model	Number of restraint violations[e]	Root mean square distance (RMSD)/residue
	Both global and local	
Stereochemical/energetic		
X-ray and NMR		
Covalent geometry	Bond lengths and angles	
Dihedral angles	phi, psi, omega, chi	
Environment measures	H-bond satisfaction; polarity matching	
X-ray		
Symmetry constraints and crystallographic packing	Close contacts	

[a]A measure of the agreement between the model and the X-ray data. [b]The degree to which an atomic position is occupied in the crystal and ranges from unoccupied (0) to fully occupied (1). [c]B-value is the crystallographic temperature factor (Debye-Waller factor). [d]For example, the number of Nuclear Overhauser Effect (NOE) distance constraints per amino acid residue. [e]The number of set experimental restraints (e.g. limits in inter-proton NOE distances) that are being violated in the final model.

'quality' of the model is therefore very important. The form this takes depends on the method used to determine the structure and several are summarized in Table II.

When all the experimental data are made available, it is possible, in principle, to calculate all these quality measures and assess the quality of the structure automatically. The assessment results could, if desired, be stored alongside the model. This would be extremely useful for future investigations based on surveys of the deposited models. Also, the possibility of the depositors themselves applying a set of agreed upon quality measures to their models before deposition, is particularly attractive. The WWW servers established by the European consortium of protein structure validation[4] at the European Bioinformatics Institute (EBI) in Europe (http://biotech.embl-ebi.uk:8400/) and at BNL in the United States provide an important first step in this direction.

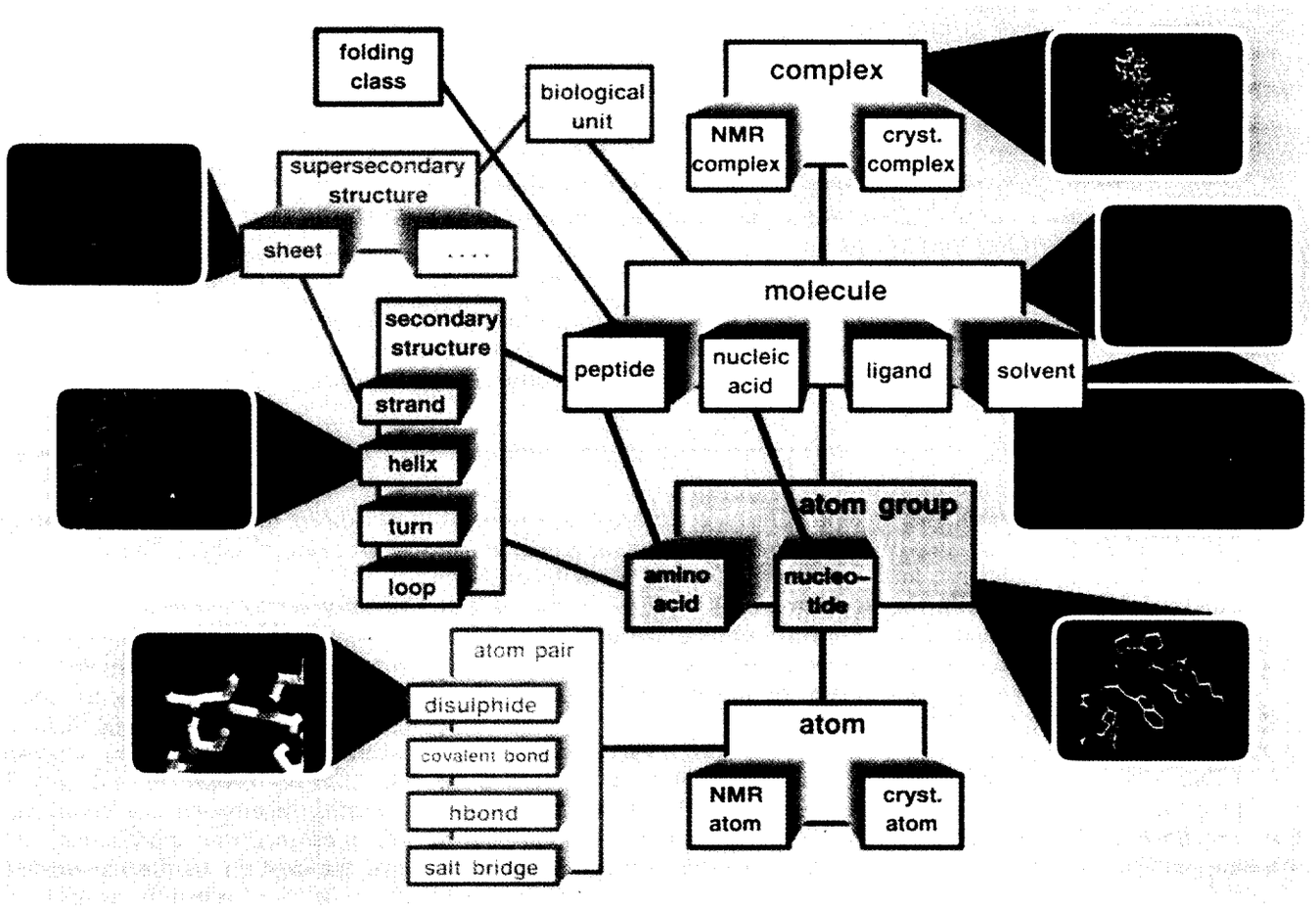

Figure 1

Database schema for macromolecular structure. Each rectangle represents a 'class' with certain properties or attributes. Relationships between classes are indicated by connecting lines. A class can be specialized into subclasses, which behave like the parent superclass, but have extra attributes or relationships, which only apply to the subclass. For example, *amino acid* is a subclass of *atom group*, which can be part of a *secondary structure*, but not of a *nucleic acid molecule*. Each subclass is shown as a named box projecting from its superclass. Existing databases (IDITIS™, SESAM) have essentially similar schemas, which are closely compatible with the mmCIF data structure.

A conceptual schema for protein structure data

ASCII files, such as those archived in the PDB, essentially store coordinates of all atoms plus information about sequence, possibly augmented with protein secondary structure and other features, but they contain nothing about the relationships between data items. When data are shared between many users and accessed by many application programs, it is vital also to capture the description of the actual objects represented by the data, the relationships between them and the constraints on their relationships and properties. This constitutes the *conceptual schema*; it is used to build the physical database to which flexible and reliable access is achieved by the database management system (DBMS). For macromolecular structure, there is an obvious hierarchy of structural objects, which lends itself rather naturally to an organization of the data into the diagram shown in Fig. 1.

Protein structure databases

A schema such as the one depicted in Fig. 1 can be implemented in a variety of database environments. There are two major kinds of DBMS in common usage, the Relational and the more recent Object-Oriented DBMS (Table I).

In relational databases the data are stored in tables, each column in the table corresponds to an attribute and each row corresponds to an instance. For example, in a protein structure database, each row in the residue table will correspond to a unique residue in the database (e.g. Asp215 in endothiopepsin), and each column will hold a residue property (e.g. secondary structure assignment). The data can be queried using the Structured Query Language (SQL), which allows associating data from different tables (an operation termed 'joining') to extract novel relationships and information.

By contrast, an object-oriented database (OODB) is organized as a network of objects, which are instances of various object classes. Instances of relationships (shown as thin arrows in Fig. 1) form a network of links between the objects, which are directly represented in an OODB. This makes complex queries that 'navigate' paths along the links to related objects much easier to express and to execute efficiently. Note that classes not only define attributes to store the data for objects, but also have attached methods that, for example, calculate derived data values on demand, or display instances in an appropriate style. Thus, the data in an object come with a ready made set of tools for manipulating it. An early prototype for this is P/FDM (Refs 5, 6), which has been used for relating sequence and structure of antibody data[7]. More recently, there is interest in object database schemas by the genome community, particularly OPM (Ref. 8; and http://gizmo.lbl.gov/opm.html), which can use relational data storage, and ACE-DB (http://probe.nalusda.gov:8000/acedocs/citeacedb.html).

In the late 1980s, several groups developed relational databases for proteins[9–12] and nucleic acids[2] (see Table I). With experience, it is clear that most biologists are not willing to master SQL to query the databases, and require a more user-friendly window-based interface, as well as tools for three-dimensional visualization. Programs such as RASMOL (Ref. 13) or MOLSCRIPT (Ref. 14) have been interfaced to a variety of databases and browsers to display structures and hit-lists.

All the available databases were initially developed as prototypes in an academic environment, but continued maintenance is often difficult and requires either commercialization as for IDITIS™ (see Table I), distributed by Oxford Molecular, or central support. The databases of derived data are already large (almost 19 gigabytes) and growing rapidly. The NDB (Ref. 2) is currently about a tenth of the size of protein database in entries, and much smaller in terms of derived data. Ideally, such large and comprehensive databases should be held and updated centrally, with access over the Internet.

However, while searches of the whole sequence database can be performed in seconds, structural queries can be much more complex and can take hours rather than seconds to complete. Thus, it is difficult to provide an open access service. Some intermediate services have been implemented through the use of menus to restrict access to simple queries and 'pre-canned' answers to frequently asked more complex queries (NDB at http://ndbserver.rutgers.edu). Establishing such services alongside the PDB would be most useful.

Box 2. Illustrative queries

Simple queries on one structure
- Select any or all of the primary information on a single protein.
- Select any or all derived information on a single protein, e.g. secondary structure, solvent accessibility, S–S bridges, salt bridges, etc.

Simple queries on all the structures
- Select all globins, or growth factors.
- Select all proteins that bind zinc (and/or nucleotides).
- Select all enzymes with a given EC number.
- Select all proteins solved by NMR spectroscopy.
- Select all proteins determined to resolution better than 1.2 Å.
- Find all dimers.
- Find all non-proline *cis*-peptide bonds.
- Find all the mainly-α proteins.

Complex queries
- Select all peptide segments of a given sequence (such as Ala-Ala-Ala) across whole database and extract specific information, e.g. solvent accessibility.
- Find the secondary structure conformation of all buried tyrosines.
- Find all salt-bridge-forming arginines in helices.
- Find all sheets with +1, +1X topology.
- Find all Type I β-turns.
- Find all aspartic acid residues that lie in a β-strand and hydrogen bond to a histidine.
- Find all Asp-His-Ser triads.
- Find all flexible regions with B-values > 50.
- Find all residues in contact with haem groups.
- Find all histidines that are protonated, i.e. donate two side chain H-bonds.

Some example queries

Database technology enables the non-expert scientist to query the protein structure data flexibly and rapidly. A major goal has been to make the querying process as simple as possible and, to this end, several groups have developed specific interfaces that allow easier access to the data[15,16]. The software company Oxford Molecular has developed a specialized window-based user-interface, which allows a 'click and extract' approach (IDITIS™). For SESAM, a simpler command-line driven interface has been built, which allows queries to be constructed from menus[13].

A list of example questions that might be asked is given in Box 2. Some of the questions are conceptually simple, but belong to the category of complex queries, because they require retrieval of information from several database tables and because some of this information represents derived data obtained from the raw atomic coordinates by applying specialized programs.

Concluding remarks

The field of macromolecular structure databases is bound to evolve dramatically as the number of known structures increases and as the scientific community becomes more widely aware

of the importance of structural data. In the future, we will probably see several centres world-wide sharing the responsibility for managing the primary archive, and a plethora of specialized databases offering access to different types of derived information.

Technological advances expected in the next few years, such as widespread use of multimedia tools, the availability of desktop multi-processor workstations combined with database systems that do not rely on spinning magnetic disks, and major improvements in bandwidth on the Internet, will radically improve our ability to handle the complex data retrieval and visualization problems that challenge today's technology. New demands, however, are coming from the nature of modern research in biology and biotechnology, where the emphasis is now on molecular genetic approaches to understanding the basis of human and animal disease. A much wider community of biologists, chemists and medical scientists now wants and needs to understand the relationships between the molecular, structural, genetic and biological function of a gene product. These scientists will ask biological databases yet more complex and searching questions and will demand fast access to very diverse information through simple and intuitive user interfaces.

The main challenges here are to archive and validate the data as soon as they become available, and the seamless integration, or cross-referencing, of genomic, sequence and structure databases. We need to meet these challenges if future science is to profit from the full potential of structural, sequence, biochemical and biological data.

Acknowledgements

The EU BRIDGE DataBase Project Consortium acknowledges the financial support from the European Community BRIDGE contract (BIOTCT910271), from the British Council, the National Fund for Scientific Research (Belgium) and the French Community of Belgium.

References

1 Bernstein, F. C. *et al.* (1977) *J. Mol. Biol.* 112, 535–542
2 Berman, H. M. *et al.* (1992) *Biophys. J.* 63, 751–759
3 Rose, M. T. (1990) *The Open Book: A Practical Perspective on OSI*, Prentice Hall
4 Wodak, S. J., Pontius, J. U., Vaguine, A. and Richelle, J. (1995) in *Making the Most of your Model* (Hunter W. N., Thornton, J. M. and Bailey, S., eds), pp. 41–51, CCL, Daresbury Laboratory
5 Gray, P. M. D., Paton, N. W., Kemp, G. J. L. and Fothergill, J. E. (1990) *Protein Eng.* 3, 235–243
6 Gray, P. M. D., Kulkarni, K. G. and Paton, N. W. (1992) *Object-Oriented Databases: a Semantic Data Model Approach*, Prentice–Hall
7 Kemp, G. J. L., Jiao, Z., Gray, P. M. D. and Fothergill, J. E. (1994) in *Applications of Databases. Proceedings of the First International Conference, ADB–94* (Litwin, W. and Risch, T., eds), pp. 317–335, Springer-Verlag
8 Chen, I. A. and Markowitz, V. M. (1995) *Information Systems* 20, 393–418, Pergamon Press
9 Morffew, A. J., Todd., S. J. P. and Snelgrove, M. J. (1983) *Comput. Chem.* 7, 9–16
10 Bryant, S. H. (1989) *Protein Struct. Funct. Genet.* 5, 233–247
11 Islam, S. A. and Sternberg, M. J. E. (1989) *Protein Eng.* 2, 431–442
12 Huysmans, M., Richelle, J. and Wodak, S. J. (1991) *Protein Struct. Funct. Genet.* 11, 59–76
13 Sayle, R. A. and Milner-White, E. J. (1995) *Trends Biochem. Sci.* 20, 374–376
14 Kraulis, P. (1991) *J. Appl. Crystallogr.* 24, 946–950
15 Boyle, J., Fothergill, J. and Gray P. M. D. (1993) in *Proceedings IEEE Visualisation '93* (Lee, J. P. and Grinstein, G. G., eds), pp. 173–185, Springer-Verlag
16 Seifert, K. L. and Rawlings, C. J. (1988) in *People and Computers* (4th edn) (Jones, D. M. and Winder, R., eds), pp. 391–406, Cambridge University Press
17 Seavey, B. R., Farr, E. A., Westler, W. M. and Markley, J. L. (1991) *J. Biomol. NMR* 1, 217–236
18 Sander, C. and Schneider, R. (1991) *Protein Struct. Funct. Genet.* 9, 56–68
19 Pascarella, S. and Argos, P. (1992) *Protein Eng.* 5, 121–137
20 Holm, L. and Sander, C. (1994) *Protein Struct. Funct. Genet.* 19, 256–268
21 Walsh, L. L. (1994) *CABIOS* 10, 551–557
22 Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995) *J. Mol. Biol.* 247, 536–540
23 Holm, L. and Sander, C. (1994) *Nucleic Acids Res.* 24, 206–210
24 Etzold, T. and Argos, P. (1993) *Comput. Appl. Biosci.* 9, 49–57
25 Peitsch, M. C., Stampf, D. R., Wells, T. N. C. and Sussman, J. L. (1995) *Trends Biochem. Sci.* 20, 82–84
26 Gerstein M., Lesk, A. and Chothia, C. (1994) *Biochemistry* 33, 6739–6749
27 Thornton, J. M., and Gardner, S. P. (1989) *Trends Biochem. Sci.* 14, 300–304
28 Shindyalov, I. N., Cooper, J., Chang, W. and Bourne, P. E. (1995) in *Proceedings of the 28th Annual Hawaii International Conference on System Sciences* (Hunter, L. and Shriver, B. D., eds), pp. 207–217, IEEE Computer Society Press
29 Shindyalov, I. N. and Bourne, P. E. (1995) *J. Appl. Crystallogr.* 28, 847–852
30 Klose, K. (1995) in *Bioinformatics and Genome Research. Proceedings of the 3rd International Conference on Bioinformatics and Genome Research* (Lim, H. A. and Cantor, C. R., eds), p. 37, World Scientific Publishing Co.
31 Barton, G. J. and Rawlings, C. J. (1990) *J. Tetrahedron Comp. Method.* 3, 739–756
32 Gilliland, G. L., Tung, M., Blakeslee, D. M. and Ladner, J. (1994) *Acta Crystallogr. D50*, 408–413
33 Bairoch, A. and Boeckmann, B. (1994) *Nucleic Acids Res.* 22, 3578–3580
34 George, D. G., Barker, W. C. and Hunt, L. T. (1986) *Nucleic Acids Res.* 14, 11–15
35 Benson, D., Lipman, D. J. and Ostell, J. (1993) *Nucleic Acids Res.* 21, 2963–2965
36 Bairoch, A. (1994) *Nucleic Acids Res.* 22, 3626–3627
37 Pattabiraman, N., Namboodiri, K., Lowrey, A. and Gaber, B. P. (1990) *Protein Seq. Data Anal.* 3, 387–405
38 Sonnhammer, E. L. L. and Kahn, D. (1994) *Protein Sci.* 3, 482–492
39 Bairoch, A. and Bucher, P. (1994) *Nucleic Acids Res.* 22, 3583–3589
40 Attwood, T. K. and Beck, M. E. (1994) *Protein Eng.* 7, 841–848
41 Henikoff, S. and Henikoff, J. G. (1994) *Genomics* 19, 97–107
42 Michie, A. D., Jones, M. L. and Attwood, T. K. (1996) *Trends Biochem. Sci.* 21, 191

CULTURE CORNER

Stranded in antisense

BZIIFNXN NZROORD

VOYRHMVH GHLN BOOZVI TMRVY HWMZIGH SGLY
GMLD H'ZMW DLOOLU
G'MLW HGHRGMVRXH GZSG
VOYRHMVUVWMR VGRFJ BZH LG GLM UR
VOYRHMVSVIKNLXMR WMZ BGRK Z H'GR

VHMVHRGMZ MR WVWMZIGH

WILLIAM C. McMURRAY
Department of Biochemistry,
University of Western Ontario,
London, Ontario, Canada.

(See p. 277 for solution)