The evolution of structural databases

Oliviero Carugo and Sándor Pongor

Starting with the Protein Data Bank (PDB) as a common ancestor, the evolution of structural databases has been driven by the rapprochement of the structural world and the practical applications. The result is an impressive number of secondary structural databases that is welcomed by structural biologists and bioinformaticians but runs the risk of producing an embarrassment of riches among non-specialist users. Given that any profit depends on the number of customers, efficient interfaces between many structural data banks must be available to make their contents easily accessible. Increasing the information content of central structural repositories might be the best way to guide users through the many, sometimes overlapping databases.

Published online: 15 October 2002

Oliviero Carugo*
Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, Area Science Park, Padriciano 99, 34012 Trieste, Italy.
Dept of General Chemistry, Pavia University, viale Taramelli 12, 27100 Pavia, Italy.
*e-mail: carugo@icgeb.trieste.it

Sándor Pongor
Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, Padriciano 99, 34012 Trieste, Italy.
pongor@icgeb.trieste.it

On one hand, molecular databases and the galaxy of linked databanks that surround them are one of the most sophisticated knowledge representation tools mankind has ever built. On the other hand, to a non-specialist they could appear as a click-able maze in which information cannot only be found but also lost.

Databases result from the interaction between three constituent parts: the models, the descriptions and the analysis.

- The models are the conceptual structures or mental representations that we use to store information on molecules.
- The formal or narrative description of the data is the backbone of the databases.
- The analysis covers everything we and our computers do with molecular data in molecular modeling, prediction, classification, similarity search, visualization and so on.

Whether we consider small molecules, proteins, metabolic pathways, genetic networks or genomes, the underlying conceptual models are reassuringly uniform. They are invariably built of entities (atoms, residues, domains, genes, etc.) and the relationships (chemical bonds, spatial or genomic proximity, etc.) between them [1,2].

In spite of this common underlying framework, the databases used in the various disciplines of biology are seemingly very different, having evolved under vastly different scientific constraints. In recent years, the data added to the sequence databases has been the most spectacular as a result of the genome sequencing efforts and the concomitant development in sequence bioinformatics. Development has been much slower for structural databases, even though the emerging structural genomic initiatives and the underlying high-throughput technologies [3] are supposed to put further emphasis on structural databases, their organization and their management. It should be noted that 3D structures provide information that is unique and crucial in many respects; atomic details, for example, are indispensable for the understanding of enzyme mechanisms [4,5]. In addition, some functionally important motifs, such as the catalytic triad of serine proteases [6], could only have been discovered from 3D structures. And last, but not least, structural biology techniques offer unprecedented possibilities to examine at a molecular level the interactions between macromolecules.

It should be noted that both structural and sequence databases have many common themes. It is customary to divide the contents of a database record into 'structure' and 'annotations' – a formal description of the molecule in terms of sequence or 3D structure, in addition to an entirely narrative description. However, annotations often contain structural information, so it is more useful to consider annotations as a list of DESCRIPTORS (see Glossary) referring either to the entire molecule or to some parts of it.

This article intends to review, from the user's point of view, a few strategic questions related to the future development of the structural databases.

Evolution of the structural databases

All structural databases have a common ancestor, the Protein Data Bank (PDB) [7], which was established in 1971 and six years later contained 77 atomic co-ordinate entries for 47 macromolecules [8]. Today, PDB contains 18 000 macromolecular structures (90% proteins, 6% nucleic acids, 4% protein–nucleic acid complexes; 82% crystal structures and 18% NMR solution structures).

Glossary

Descriptors: add information to a defined range of database items. For example, descriptors can refer to entire proteins (e.g. name), their parts (name of a domain) as well as to groups of proteins (protein families). Biological descriptors that could increase the use of structural databases include the definition and classification of the biological function, phylogenetic origin, subcellular location, role in disease and so on. Descriptors derived from the raw structural data include the protein architecture as well as its simplified definitions, fold types, quaternary structure (especially when ambiguously defined by the crystallographic symmetry). Molecular engineering details must also be described, in order distinguish a full-length construct from a smaller fragment, a native molecule from a mutant, or native isoforms.

Ontology: A formal definition of concepts (entities, relationships) of a given area of knowledge, described in a standardized form. In biology, ontologies developed for the genes and proteins of given organisms include metabolic, genetic and product-interaction networks. The current infrastructure of PDB includes a comprehensive ontology for macromolecular structure and experiment.

 0167-7799/02/$ – see front matter PII: S0167-7799(02)02082-6

Box 1. Websites of some structural databases

Primary resource	
Protein Data Bank	http://www.rcsb.org
Information related to the primary resource	
Macromolecular Structure Database	http://www.ebi.ac.uk/msd/index.html
Nucleic Acid Database Project	http://ndbserver.rutgers.edu/NDB/index.html
BioMagResBank	http://www.bmrb.wisc.edu/Welcome.html
Protein domain and fold databases	
3Dee	http://jura.ebi.ac.uk:8080/3Dee/help/help_intro.html
CATH	http://www.biochem.ucl.ac.uk/bsm/cath
HSSP	http://www.sander.ebi.ac.uk/hssp
SCOP	http://scop.mrc-lmb.cam.uk.ac/scop
Examples of specialized resources	
BIND – binding database	http://www.bind.ca/index.phtml?page=databases
BindingDB – binding database	http://www.bindingdb.org/bind/index.jsp
Decoys 'R' Us	http://dd.stanford.edu
Disordered structures	http://bonsai.ims.u-tokyo. ac.jp/~klsim/database.html
DNA binding proteins	http://ndbserver.rutgers.edu/structure-finder/dnabind/
Intramolecular movements	http://molmovdb.mbb.yale.edu/MolMovDB/
Loops	http://www-cryst.bioc.cam.ac.uk/~sloop/Info.html
Membrane protein structures	http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html
Metal cations	http://metallo.scripps.edu/
P450 containing systems	http://www.icgeb.trieste.it/~p450srv/
Predicted protein models	http://guitar.rockefeller.edu/modbase
Protein-DNA contacts	http://www.biochem.ucl.ac.uk/bsm/DNA/server/
Protein–protein interfaces	http://www.biochem.ucl.ac.uk/bsm/PP/server/
ProTherm	http://www.rtc.riken.go.jp/jouhou/protherm/protherm.html
Quaternary structure	http://pqs.ebi.ac.uk
Small ligands	http://alpha2.bmc.uu.se/hicup/
Small ligands	http://www.ebi.ac.uk/msd-srv/chempdb
The Protein Kinase Resource	http://pkr.sdsc.edu/html/index.shtml
Examples of search/retrieval facilities and database interfaces	
3DinSight – structure and function of biomolecules	http://www.rtc.riken.go.jp/jouhou/3dinsight/3DinSight.html
BioMolQuest – structure and function of proteins	http://bioinformatics.danforthcenter.org/yury/public/home.html
Entrez	http://www3.ncbi.nlm.nih.gov/entrez/query.fcgi
Image Library of Biological Micromolecules	http://www.imb-jena.de/IMAGE.html
OCA	http://bioinfo.weizmann.ac.il:8500/oca-docs/
PDBSUM	http://www.biochem.ucl.ac.uk/bsm/pdbsum/
ProNIT – protein nucleic acid interactions	http://www.rtc.riken.go.jp/jouhou/pronit/pronit.html
TargetDB	http://targetdb.pdb.org/
SRS	http://srs.ebi.ac.uk/

In addition to the quantity, the phenotype and the quality of the deposited structures have also changed. Owing to new experimental techniques (e.g. PCR, synchrotrons and high-field NMR resources) the quality of data has steadily improved and larger and larger macromolecular structures have been determined. The format of the PDB records has also adapted to accommodate these changes. Since its beginning, PDB represented 3D structures with two types of records, one with the atomic co-ordinates and the other with annotation information (name of the molecule, sequences etc.). The number of records not related to the atomic positions increased from ~200 in 1977 to ~700 in 2002. However, most of the additional information refers to experimental or computational technicalities rather than to genuine biological features. The PDB was created as a crystallographic database and, despite the fast growing body of NMR data, this remains its legacy. This limitation, observed long ago [8], depends on the fact that the majority of users accessing PDB today are not crystallographers or structural biologists but rather biochemists and molecular biologists.

Over the years, a conspicuous number of homologous databases have evolved from the PDB (see Box 1). Many of them concentrate on various classes of structural features, such as protein domains [9–12], loops [13], contact surfaces [14,15], quaternary structure [16], small-molecule ligands [17], metals [18] and disordered regions [19]. Other databases concentrate on biological themes. Databases of membrane proteins [20] or of selected protein families, such as kinases [21] or P450 containing systems [22], are typical examples of adding biological information to structural data. The Protherm database contains thermodynamic and kinetic data, linked to protein sequence and structure databases [23]. The BIND and BindingDB databases list experimental data on macromolecular binding [24–26]. Some of the more distantly related databases

contain specific structural data such as the collection of experimentally determined intra-molecular movements [27] or theoretical models that can be used to test novel theoretical prediction protocols [28].

Some databases have an obvious, enormous impact on molecular biology. For example, the domain and fold collections (3Dee [12], CATH [10], HSSP [9] and SCOP [11]) list and classify protein 3D domains differently. Listing and classification is essential especially for multi-domain macromolecules and is extremely important for structure and function prediction as well as in evolutionary studies. Other secondary databases concentrate on highly specific fields and their potential audience is consequently rather narrow.

The consequence of all these efforts is the development of new content, which consists of adding additional information to subsets of structures or structural moieties. However, it is relatively quiet on the technology front. Most of the collections are maintained and developed in the form of relational databases, whereas the distribution of the databases is usually still in the flat-file format [29].

Refolding the unfolded data

The process that created a plethora of structural databases is, in a way, similar to the unfolding of a protein. Given that the unfolded molecules usually cannot carry out the function of the native protein, similarly the multitudes of secondary databases do not offer either a comprehensive picture or easy access to information. Database diversity has over-reached itself and offers little additional benefits to those users whose needs originally created it.

Can structural databases be refolded? And why they should refold? The answer is closely related to a more general problem of standardizing scientific information [30] and modeling protocols [31]. An important step towards database inter-operability is the standardization of the conceptual framework into database ONTOLOGIES [32]. In principle, integration seems to be the only avenue that can lead us out of the trap of database diversity.

Creating a unique super-database for all biological data would be surely too expensive in terms of personnel costs. A less ambitious alternative proposes the use a battery of federated databases that are queried over the network through a common protocol [29]. This approach would allow distributed database maintenance – a crucial point considering both cost and expertise – the first practical applications are yet to appear.

Another possibility is to access locally maintained copies of several databases through a common interface. There are some overheads but these seem to be worthwhile for non-specialist users of biological databases and especially for non-academic users concerned with proprietary information.

Some justly popular data-retrieval tools, such as SRS [33] and DBGET/LinkDB [34], are able to manage an impressive number of different data, from bibliographic to structural databases. The user can search several databases simultaneously and obtain a list of database entries. Other steps in improving the interfaces between databases have been done. Entrez [35], for example, allows one to search for macromolecular 3D structures, sequences and related bibliographic information but does not interface with most of the structural databases. Genetic, functional, phylum or disease related information is interfaced to the PDB by the OCA browsing system at the Weizmann institute. The PDB search facility itself can be customized to include not only structural search restraints but also biological information (e.g. carbohydrates can be discriminated from enzymes and DNA). Collections of links to several databases are provided by PDBSUM [36] and by the Jena Image Library of Biological Macromolecules [37]. TargetDB is an interesting example of voluntary collaboration for maintaining a structural genomics database (http://targetdb.pdb.org/).

A frequent limitation of the current multi-database retrieval systems is that they usually offer only click-able links, which can be impractical when trying to evaluate a large number of structures. However, the easy-to-follow links are very popular among the non-specialist users and so it is worthwhile to consider a broadening of the services of PDB along the lines of the development seen with sequence databases.

Several integrated relational databases have also been developed recently [23,38–40]. In general, they do not use the original flat files of different databases but integrate the information into a single 'super database', usually defined a data warehouse, which can be managed through commercial or open-source databases management systems. This approach allows users to make flexible searches by combining various keywords and conditions using the Structures Query Language (SQL). For example, BioMolQuest [38] integrates PDB (3D structures) [7], CATH (definition and classification of folding domains) [10], SWISS-PROT (protein sequences and some functional information) [41] and ENZYME (protein functions) [42]. Database integration is a formidable task because of the variability in language and format and because of the numerous minor inconsistencies between different databases and even within the same database. For instance, An *et al.* [39] pointed out that the protein sequence information is reported twice, sometimes inconsistently, in each PDB file.

Users who are interested in the atomic level of protein structures constitute a large critical mass representing such broad fields as structural biology, drug design and so on. This large user community now often uses the more easily accessible and well cross-referenced sequence databases as the starting point, even though they would probably prefer to start from the structural motifs they are actually working on. When you consider the numerous structural genomics efforts that mirror the genomic sequence

determinations, PDB has an excellent possibility of becoming the central service of a fast growing user community, especially within the academic world.

A re-shaping of PDB is already underway [43] but further improvement could be suggested, for example through minor modifications to the PDB lexicon. For example, PDB has a tradition to distinguish ATOM or HETATM fields – the first refers to protein or DNA atoms, the second to any other molecule present in the structure. As a consequence, post-translational modifications, such as phosphorylated amino acids, cofactors, enzyme inhibitors or glycans and such like are, in general, not directly recognized by molecular modeling or analysis software. Using additional labels other than ATOM and HETATM would greatly facilitate the handling of structures containing various types of molecules, whereas HETATM could be retained for compounds without a clear biological relevance.

Several additional descriptors might also be considered and inserted into the PDB flat files. For example, the phylogeny of the organism and the cellular location and functional description of the protein are worthy of inclusion, even at the risk of repetition. Substructure descriptors, such as active sites, subunit interfaces or domains and motifs assigned within secondary databases (as well as their cross references to domain sequence collections) could be added in the form of a feature table. Many PDB entries contain protein assemblies but a comment on whether or not this is a crystallization artifact or a functional assembly, is often missing.

The suggestions listed here of course reflect subjective opinions. However, it would be extremely useful to incorporate into the PDB files the most relevant secondary structural databases, making PDB an integrated warehouse. Nevertheless, the file structure should be very flexible to easily accommodate new features that could appear in the future.

Finally, an important difference between the PDB and the many secondary structural databases should be noted. Although PDB resulted from a multi-laboratory, international effort, most of the secondary databases were produced by single laboratories. It is therefore important to ask ourselves if a more open and cooperative approach could be applied for maintaining and refolding structural databases.

Acknowledgements
This work was partly funded by the EU-project ORIEL (IST-2001–32688) 'On-line research environment for the life sciences', coordinated by the European Molecular Biology Organization (EMBO).

References

1 Pongor, S. (1988) Novel databases for molecular biology. *Nature* 332, 24
2 Hatsagi, Z. *et al.* (1994) Protein motifs: Towards a unified view of databases. *Proceedings 27th Anuual Hawaii International Conference on System Science* 255–264
3 Chance, M.R. *et al.* (2002) Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci.* 11, 723–738
4 Lamzin, V.S. *et al.* (1995) How nature deals with stereoisomers. *Curr. Opin. Struct. Biol.* 5, 830–836
5 Dauter, Z. *et al.* (1995) Proteins at atomic resolution. *Curr. Opin. Struct. Biol.* 5, 784–790
6 Blow, D. (1997) The tortuous story of Asp.His.Ser: structural analysis of alpha-chymotrypsin. *Trends Biochem. Sci.* 22, 405–408
7 Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242
8 Bernstein, F.C. *et al.* (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112, 535–542
9 Sander, C. and Schneider, R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9, 56–68
10 Orengo, C.A. *et al.* (1997) CATH-a hierarchic classification of protein domain structures. *Structure* 5, 1093–1108
11 Murzin, A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540
12 Siddiqui, A.S. *et al.* (2001) 3Dee: A database of protein structural domains. *Bioinformatics* 17, 200–201
13 Donate, L.E. *et al.* (1996) Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. *Protein Sci.* 5, 2600–2616
14 Jones, S. and Thornton, J.M. (1996) Principles of protein–protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* 93, 13–20
15 Luscombe, N.M. *et al.* (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.* 1, REVIEWS001
16 Henrick, K. and Thornton, J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* 23, 358–361
17 Kleywegt, G. and Jones, T. (1998) Databases in protein crystallography. *Acta Crystallogr. D* 54, 1119–1131
18 Castagnetto, J.M. *et al.* (2002) MDB: the metalloprotein database and browser at the Scripps Research Institute. *Nucleic Acids Res.* 30, 379–382
19 Sim, K.L. *et al.* (2001) ProDDO: a database of disordered proteins from the Protein Data Bank (PDB). *Bioinformatics* 17, 379–380
20 White, S.H. and Wimley, W.C. (1999) Membrane protein folding and stability: Physical priciples. *Annu. Rev. Biophys. Biomol. Struct.* 28, 319–365
21 Smith, C.M. *et al.* (1997) The protein kinase resource. *Trends Biochem. Sci.* 22, 444–446
22 Degtyarenko, K.N. (1995) Structural domains of P450-containing monooxygenase systems. *Protein Eng.* 8, 737–747
23 Gromiha, M.M. *et al.* (2000) ProTherm, version 2.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* 28, 283–285
24 Salama, J.J. *et al.* (2001) Automatic ammotation of BIND molecular interactions from three-dimensional structures. *Biopolymers* 61, 111–120
25 Bader, G.D. and Hoguew, C.W. (2000) BIND–a data specification for sorting and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* 16, 465–477
26 Chen, X. *et al.* (2001) The binding database: overview and user's guide. *Biopolymers* 61, 127–141
27 Gerstein, M. and Krebs, W. (1998) A database of macromolecular motions. *Nucleic Acids Res.* 26, 4280–4290
28 Samudrala, R. and Levitt, M. (2000) Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. *Protein Sci.* 9, 1399–1401
29 Valencia, A. (2002) 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? *EMBO Rep.* 3, 396–400
30 Grivell, L. (2002) Mining the bibliome: searching for a needle in a haystack? *EMBO Rep.* 3, 200–203
31 Kitano, H. (2002) Standard for modelling. *Nat. Biotechnol.* 20, 337
32 Westbrook, J.D. and Bourne, P.E. (2000) STAR/mmCIF: an ontology for macromolecular structure. *Bioinformatics* 16, 159–168
33 Zdobnov, E.M. *et al.* (2002) The EBI SRS server – new features. *Bioinformatics* 18, 1149–1150
34 Fujibuchi, W. *et al.* (1998) DBGET/LinkDB: an integrated database retrieval system. *Pac. Symp. Biocomput.* 3, 681–692
35 Wheeler, D.L. *et al.* (2002) Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.* 30, 13–16
36 Laskowski, R.A. (2001) PDBSUM: summaries and analyses of PDB structures. *Nucleic Acids Res.* 29, 221–222
37 Reichert, J. and Suhnel, J. (2002) The IMB Jena Image Library of Biological Macromolecules: 2002 update. *Nucleic Acids Res.* 30, 253–254
38 Bukhman, Y.V. and Skolnick, J. (2001) BioMolQuest: integrated database-based retrieval of protein structural and functional information. *Bioinformatics* 17, 468–478
39 An, J. *et al.* (1998) 3DinSight: an integrated relational database and search tool for the structure, function and properties of biomolecules. *Bioinformatics* 14, 188–195
40 Prabakaran, P. *et al.* (2001) Thermodynamic database for protein-nucleic acid interactions (ProNIT). *Bioinformatics* 17, 1027–1034
41 Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence databse and its supplement TrEMBL. *Nucleic Acids Res.* 28, 45–48
42 Bairoch, A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.* 28, 304–305
43 Westbrook, J. *et al.* (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.* 30, 245–248