Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Protein structure resources

Helge Weissig[a] and Philip E. Bourne[a,b]*

[a]San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0505, USA, and [b]Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Correspondence e-mail: bourne@sdsc.edu

Received 1 January 2002
Accepted 28 February 2002

The Protein Data Bank (PDB) is the primary source of macromolecular structure data for a worldwide community of users. A subset of those users then process these data to derive secondary information which is also available on the WWW. This process includes validation, some form of reductionism, *via* sequence or structure, or visualization. The result, a set of further web-accessible resources on protein structure and functional classification, links to primary genomic information, protein–protein and protein–ligand interactions, protein dynamics and protein-modeling resources. This paper reports on these processes and a subset of the web resources that result.

1. Introduction

The structural biology community is unique in that for over 30 years there has only been one repository for primary macromolecular structure data. That repository is the Protein Data Bank (PDB; http://www.pdb.org; Bernstein *et al.*, 1977; Berman *et al.*, 2000, 2002). From the viewpoint of data movement, all developments in structural biology have stemmed from an outward flow of data from the PDB (Fig. 1) to a large number of resources and individual research laboratories, a process that was accelerated by the advent of the World Wide Web. The name *Protein* Data Bank is somewhat misleading, as the repository contains data on *all* biological macromolecules, *i.e.* proteins *and* DNA, RNA, carbohydrates and all complexes thereof. This article focuses on the outward flow of protein structure data as found in the PDB and how those data are being used.

Each week, at the request of depositors, the PDB releases new curated and annotated data to the community, presently of the order of 30–100 structures per week. Thus, the PDB is providing the *primary* data from which much *secondary* information is derived. Stated another way, taking the primary PDB data, many researchers worldwide perform a set of *actions* on that data, such as validation, classification, provision of non-redundant sequence and structure sets (reduction), visualization and so on (Fig. 1). In short, they add value which when placed in *resources* accessible to the community adds a great deal to the field of structural biology and hence permits a better understanding of how living systems function and how to treat disease states. There are many types of secondary resource. We have chosen a few which are used widely and are relevant to this readership. They are resources that (i) relate to specific families of proteins, (ii) relate structure to the genomes from which they were transcribed, (iii) study protein–protein interactions or (iv) study protein motions and protein modeling (Fig. 1).

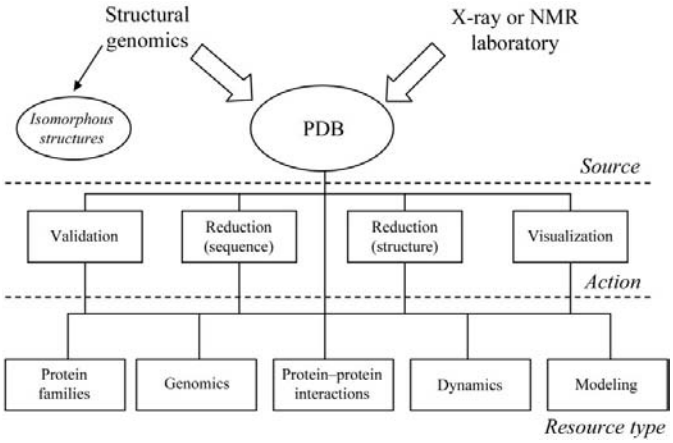

Figure 1
The flow of macromolecular structure data. Macromolecular structure data can be thought of as existing at three levels. *Source* is the primary data. At present these data come only from the PDB, but if partial and/or incomplete structures are released directly by structural genomics projects there will be a new source of primary data. *Action* implies the processing of these primary data in some way to provide derived information. *Resources* consist of various Internet-accessible resources that present these data to the community.

1.1. Primary PDB data

Primary data are defined here as data derived directly by experiment using the techniques of X-ray crystallography and NMR[1] and includes the name of the macromolecule and all its components, the source of the macromolecule and all its components, the primary sequence and chemical formula of all components and the atomic Cartesian coordinates. Over the years, the PDB has come to be regarded as the repository for the final data set on a given structure, even though many substitutions have been made (http://pdbobs.sdsc.edu; Weissig & Bourne, 1999) by depositors providing updated data sets. Scientists expect to come to the PDB to obtain the best and most complete source of primary data on the structure of biological macromolecules.

To assist in uniquely identifying a structure, each one is provided with a PDB identifier (PDBid) of the form [(0–9)(*a*–*z*, 0–9)(*a*–*z*, 0–9)(*a*–*z*, 0–9)]; for example, 4hhb. The PDBid is then the immutable reference to that structure. In other words any reference to, for example, 4hhb in the published literature or another database references *exactly the same* data set. This is clearly critical if science is to proceed in an ordered way. The PDB staff can make minor changes to the annotation supplied with that entry without changing its PDBid. Such changes are documented in the PDB file by means of a REVDAT record. Any major change, say to the atomic coordinates, must be supplied by the original author and it would result in an existing PDBid being superseded by a new PDBid. This relationship is defined in both the old and the new PDB file. Clearly, the original file must always be available since a direct relationship between a literature reference and the data described must be maintained, even if users are encouraged to use the most recent and presumably more accurate data on a given structure. All the secondary resources described here make reference to the original PDBid and all derive their information from the same primary data.

[1] The PDB also contains a small number of theoretical models. These are mostly discarded by secondary resources using PDB data and are not considered further here.

1.2. The PDB web site *versus* the PDB file

It is important to understand the concept of the PDB web site *versus* the PDB file. The PDB files are distributed on a weekly basis *via* the PDB's ftp site at ftp://ftp.rcsb.org/pub/pdb/ and contain only the primary data. The web site provides additional tools and resources for access to these data and includes limited secondary data. Most importantly, the PDB acts as an Internet portal (gateway) to data derived from structure that are important to understanding the underlying biology. This is performed in one of three ways.

(i) Provision of a current set of web hyperlinks to relevant derived data resources. This is reviewed and maintained manually by PDB staff (see http://www.rcsb.org/pdb/links.html). All the resources described in this article are listed here.

(ii) Access to a list of the most common and well tested methods to derive secondary data. Consider the issue of structure comparison, also called structure neighboring. Structure-comparison methods give somewhat different results based on, among other things, the different underlying heuristics used to make the problem computationally tractable. The PDB does not select a specific method to impose on the user, but rather conveniently suggests a variety of well tested methods to explore.

(iii) Provision of a current set of web hyperlinks to each structure in the PDB. This is updated periodically and is maintained automatically using software known as the *Molecular Information Agent* (*MIA*; http://mia.sdsc.edu). These links include the resources discussed in this article and illustrate the many ways in which the scientific community exploits macromolecular structure data. It is worth understanding how these links are derived to illustrate the richness of the available secondary sources of macromolecular structure-related information.

1.3. The *Molecular Information Agent* (*MIA*)

MIA can be thought of as a web crawler specific to molecular biology that sniffs out useful hyperlinks to relevant derived information on a given structure that is faithfully maintained on a variety of web sites worldwide. *MIA*, originally developed in the laboratory of Michael Gribskov at the University of California, San Diego and developed further by the PDB, automatically seeks information from approximately 60 different manually chosen web resources of secondary-structure information available worldwide. As such, it provides a good measure of the depth of derived information available for any structure in the PDB. When a structure is released by the PDB clearly there is no secondary

information immediately available from other resources. However, within a brief period of one week or so the PDB file for that structure will have been downloaded to a series of web resources, processed and derived information provided as described in Fig. 1. In other words, this primary information has percolated down to a series of secondary resources that add value to the original data and make it available on the web. *MIA* is used to automatically query those resources using the PDBid as a query term and if derived information is found a synopsis of that information is returned to the PDB and stored in one of the PDB's databases along with the link back to the web resource from which it was derived. This forms the 'Other Sources' information that the PDB provides on each structure from its web site (http://www.pdb.org). An example of an 'Other Sources' page is given in Fig. 2 for human deoxyhemoglobin. The process is, in a limited way, iterative. That is, information returned from a given resource can be used to form a further query. Clearly, this process must be restricted, or a tangled web of links will result. What is presented to the user by *MIA* is (i) a set of direct links – the PDBid has been found at another resource, and (ii) indirect links – other PDBids have been found that are related to the PDBid of the original structure, usually through sequence homology or a recognized functional relationship documented at the remote resource. *MIA* is periodically run for each structure in the PDB to generate a new and updated set of links. It is a discussion of the breadth and characteristics of the resources that *MIA* searches that form the remainder of this article and provides an overview to the wealth of information derived from structure that is available to a worldwide community.

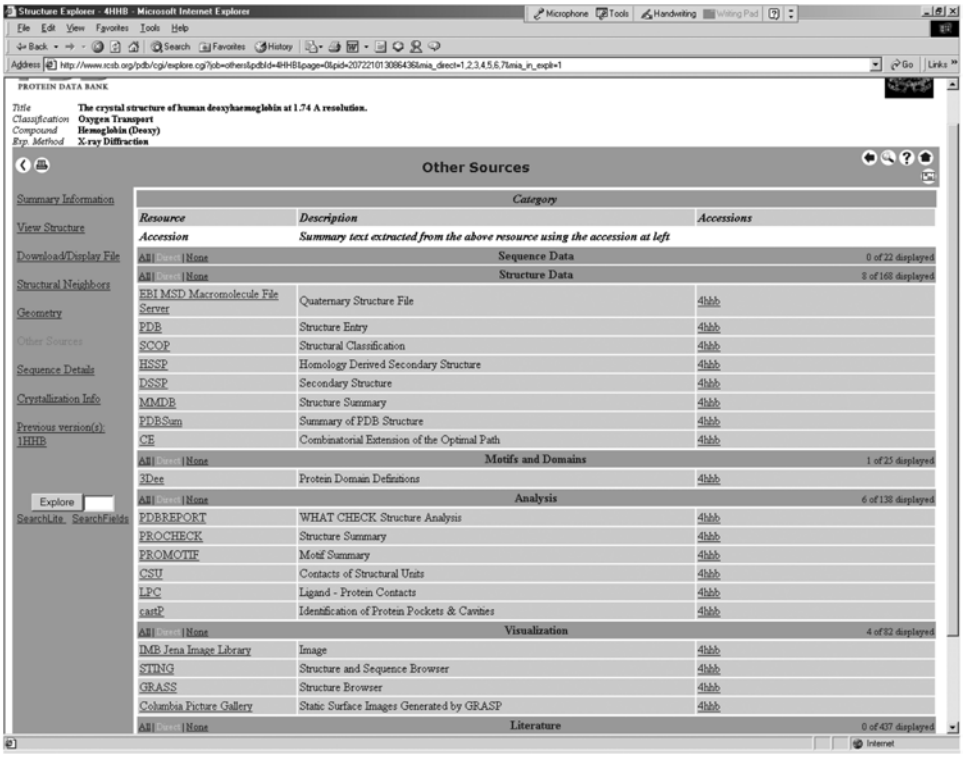

Figure 2
Example of the PDB's 'Other Sources' web page for deoxyhemoglobin (PDB code 4hhb).

1.4. Uniquely identifying secondary data

It is important to understand how secondary sources of structural information uniquely identify their contents – many do more than reference the PDBid. Different uses of structure data require that the data be thought of in different chunks (units). For example, resources that validate structure information require that the whole structure be validated as represented by the PDBid. Resources related to the primary protein or DNA sequence deal with units of *polypeptide chains* or *DNA strands*. Within each compound there are one (monomeric) or more chains and/or strands (polymeric), each of which has a unique single character identifier. For example the *A* chain of hemoglobin is often referred to as 4hhb:A (or 4hhb_A). Thus, for example, a polymeric structure while having a single PDBid is mapped to two or more entries in the sequence databases. Each reference back to the original structure in the sequence database will include the PDBid and the chain identifier.

Function is usually made with reference to *domains* – compact folding units within a protein, part or all of one or more polypeptide chains that have a recognized biological function; for example, an ATP-binding domain. There is no common nomenclature for domains, but resources using them will always reference the PDBid and chain identifier. Finally *ligands*, either covalently or non-covalently bound to the protein or other macromolecule, may be the major point of interest. Ligands in the PDB are uniquely identified by a three-letter code that in turn references a dictionary entry with full systematic name and chemical connectivity (see http://www.rcsb.org/pdb/info.html#File_Formats_and_Standards).

While these are the most common forms of protein structure classification at the coarsest level, other classifications are frequently made and relate, in part, to the notion that protein structure space is a continuum rather than discrete. Readers wishing to consider this notion should consult Shindyalov & Bourne (2000) and Holm & Sander (1996). Notwithstanding, the important issue here is that the secondary resources discussed use these compounds, chains/strands, domains and ligands as comparative reference points, each with a standard nomenclature.

Table 1
Popular software and resources for protein structure validation.

Resource	Details
PDBSum	Summaries for all protein structures including validation checks, http://www.biochem.ucl.ac.uk/bsm/pdbsum/
PROCHECK	Structure-validation suite, http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html (Laskowski *et al.*, 1997).
WHAT_CHECK	Detailed stereochemical quality summaries for all protein structures. Part of the *WHATIF* package, http://www.cmbi.kun.nl/gv/whatcheck/
SFCHECK	Validates the experimental structure factors associated with an X-ray diffraction experiment (Vaguine *et al.*, 1999).
PDB Validation Server	Validates the format and content of a PDB entry using the same software procedures as used by the PDB. Includes those listed above in this table. http://pdb.rutgers.edu/validate/.
Protein–protein interaction server	http://www.biochem.ucl.ac.uk/bsm/PP/server/server_help.html (Jones & Thornton, 1996)
Protein–DNA interaction server	http://www.biochem.ucl.ac.uk/bsm/DNA/server/ (Jones *et al.*, 1999)

2. Secondary protein structure resources

This section considers some of the *actions* performed on structure data obtained from the PDB which subsequently leads to a network of secondary data *resources* of biologically relevant data derived from the primary PDB data (Fig. 1). It is not possible to be comprehensive and brief in this discussion. The intent here is to provide an overview of what is available and for each class of resource provide suitable references and web sites for further review if desired. What should be clear is that the PDB is a very rich source of derived information from which structural biology has gained immensely. We begin with actions (validation, reduction and visualization) which are commonly applied to primary PDB data.

2.1. Validation

Here, validation implies the action of comparing structure data against well defined standards and reporting on the consistency of these data against known physical parameters and other data known to be of high accuracy; for example, the stereochemistry of small-molecule structures of amino acids and nucleic acids determined at very high resolution. Since macromolecular structures are complex, many items can be validated, but most frequently this implies the stereochemical quality of the structure. As the number of structures in the PDB increased, it became apparent that they were not uniform in their quality. Clearly, much of this lack of uniformity has to do with the nature of the experiment – data are obtained at different resolutions, have different amounts of intrinsic disorder (sometimes a functional requirement) and so on. However, it was also clear that data were not being reported systematically (Weissig & Bourne, 1999) and mistakes were being made (Morris *et al.*, 1992; Kleywegt & Jones, 1995; Hooft *et al.*, 1997) that were not being detected by the PDB. If not addressed, these problems impact significantly the value of structural data. Fortunately, the community rallied to the cause by providing comprehensive tools for the validation of both protein structures and nucleic acids. Today, these programs are used by the PDB when a structure is submitted. Popular validation software and the associated servers are outlined in Table 1.

2.2. Reduction

Structural biology is now primary-data rich, with over 17 000 structures in the PDB. To understand the scope of our knowledge of structure it is often necessary to cluster the data we have in order to better interpret it. That is, reduce what we have to deal with by looking for commonalities in the data. However, with a complex relationship between sequence and structure – structure is more evolutionarily conserved than sequence – both structure and sequence clustering is needed to cover what needs to be studied. Both structure and sequence clustering are discussed here.

2.2.1. Reduction: structure (structure classification). As early as 1980, with less than 100 macromolecular structures known, efforts were made to classify protein structures (Lesk & Chothia, 1980), since it was observed that structure was conserved even at low sequence identities and across apparently different biological functions. Structure provided a new and valuable method of reductionism that provides information not available from sequence and functional (*e.g.* EC numbers) classification schemes. As the number of structures increased, this need became even more profound in order to understand the relationship between structure and biological function. Today, we have a number of structure-classification schemes for both proteins, DNA and RNA that embody different methodologies and philosophies of what is important and should be highlighted. Table 2 summarizes popular resources that structurally classify proteins and which are characterized by being current with respect to the PDB and widely used by the community. For many biologists, these derived resources are the first point of contact with macromolecular structure, since they provide a level of organization not provided by the PDB. At the heart of these resources are differing structure-comparison methods: by eye, by algorithm and by a combination. Most (CE is the exception) first classify the structure into domains and then classify common domains. CE classifies complete polypeptide chains. While there is close agreement at the coarsest level of classification based on secondary structure – all-α, $\alpha + \beta$, α/β, all-β and random coil – departures soon arise, since schemes differ in how they define domains and indeed what and how domains are aligned, a reflection of the continuity of protein fold space. We restrict ourselves to a brief discussion highlighting features of SCOP and CATH, the most popular resources, to illustrate structure-classification schemes. Readers are referred to the papers on each resource shown in Table 2.

SCOP attempts to capture both structural and evolutionary relatedness and as such the principal divisions are family,

Table 2
Resources classifying protein structure.

Resource	Details
SCOP	The structure classification of proteins, http://scop.mrc-lmb.cam.ac.uk/scop/ (Murzin *et al.*, 1995).
CATH	Class (C), architecture (A), topology (T) and homologous superfamily (H), http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html (Orengo *et al.*, 1997).
DALI	DALI domain dictionary, http://www.embl-ebi.ac.uk/dali/domain/ (Dietmann & Holm, 2001).
VAST	Vector alignment search tool, http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml (Gibrat *et al.*, 1997).
CE	Polypeptide-chain comparison, http://cl.sdsc.edu/ce.html (Shindyalov & Bourne, 1998).
3Dee	Protein domain definitions, http://jura.ebi.ac.uk:8080/3Dee/help/help_intro.html (Siddiqui & Barton, 1995).
CAMPASS	Cambridge database of protein alignments organized as structural superfamilies, http://www-cryst.bioc.cam.ac.uk/~campass/ (Sowdhamini *et al.*, 1998).

superfamily and fold. Family implies a clear evolutionary relationship which without other evidence implies a sequence identity of 30% or greater. However, since proteins which are known experimentally to belong to the same family may have significantly less than 30% sequence identity, other rules are applied to define the family relationship. An example here would be the globins. A superfamily implies proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable. For example, actin, the ATPase domain of the heat-shock protein and hexakinase together form a superfamily. Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. Different proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. In some cases, these differing peripheral regions may comprise half the structure. Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies.

CATH also attempts to capture both structural and evolutionary relatedness. Class (C) defines the secondary structure content, architecture (A) captures the arrangement of secondary structures that historically became recognized as repetitive and were given characteristic names such as Greek-key motif and jelly roll. Topology (T) defines the overall arrangement of secondary structures. Homology (H) defines sequence homologs sharing a common evolutionary ancestor. H is further classified according to sequence (S) based on the level of sequence identity.

2.2.2. Reduction: sequence. The PDB as the complete corpus for all macromolecular structures is highly redundant with respect to sequence: this arises in the following ways. First, different groups can determine essentially the same structure independently. Secondly, a post-translational modification leading to a point mutation is almost identical in sequence. Depending on what a user is seeking, sequence redundancy can be a blessing or a curse.

The PDB provides sequence reduction as part of its search strategy and the method is described to illustrate a group of related approaches to the problem of sequence redundancy. Once computed, these methods provide a representative sequence for a group of related structures. Clearly, the representative should be the 'best' of the group. The following describes one procedure for selecting the best. Sequence homologs are clustered according to an algorithm developed in the laboratory of Adam Godzik (Li *et al.*, 2001). Clustering is applied to all protein chains of at least 20 amino acids. Three sets of clusters are generated with 90, 70 and 50% sequence identity. In each cluster, the chains are sorted (*i.e.* ranked) according to the following criteria (in the following order).

(i) Experimental structures *versus* theoretical models (models are given the lowest ranks).

(ii) A simple quality factor, calculated as 1/resolution $-$ R value.

(iii) Deposition date (newer structures have higher ranks).

(iv) Alphabetically.

This ranking has the following implications.

(*a*) Higher quality structures (better resolution, lower R value) are preferred.

(*b*) Structures determined by X-ray crystallography are preferred over NMR structures.

(*c*) A theoretical model will only represent itself (or other models).

The selection of representative structures based on the clusters of chains is performed as follows.

(I) All structures that **do not** contain at least one protein chain of 20 or more amino acids will automatically represent themselves.

(II) All structures that **do** contain at least one protein chain of 20 or more amino acids are processed as follows.

(*a*) Generate a list of all protein chains of 20 or more amino acids in the set of structures.

(*b*) Obtain the cluster number and rank number for each chain.

(*c*) From each cluster number, pick the chain with the highest rank number. This comprises a non-redundant set of chains.

(*d*) Return every PDBid present in this non-redundant set of chains.

(III) The combined set of structures from (I) and (II) is returned as the selected set of structures.

In this way, a query or the results of a previous query can be reduced to a non-redundant set based on levels of sequence identity and the 'best' structure reported. The approach described here is a variation of that used by Hobohm & Sander (1994) to provide the PDBselect set of structures. Brenner *et al.* (2000) have developed the Astral resource (http://astral.stanford.edu/), again in the same spirit of reduction based on sequence, but starting from the SCOP domain classifications.

2.2.3. Visualization. In terms of primary data, a macromolecular structure is nothing more than a list of Cartesian coordinates. It is not possible for a human to ascertain the biological significance of a structure by reviewing a tabular list

of numbers. In one sense, molecular graphics provides another form of reductionism, turning that indecipherable table into a three-dimensional view that is easily interpreted. Richardson (1985) and others pioneered the visual depiction of protein structure, first with hand drawings and later with computer graphics (Levinthal *et al.*, 1968). While there are many tools for creating molecular visualizations (Tate, 2002), there are resources that provide predefined views of molecular structures. The goal of these resources is to capture the knowledge that the author has about the structure in the form of an image. So for example, calmodulin, a calcium-binding protein, should clearly indicate how the calcium is bound by the classic EF-hand motif.

An important distinction needs to be made when reviewing molecular visualizations. For X-ray crystal structures, the PDB contains the content of the asymmetric unit of the crystal. This may not be the biologically relevant entity and hence a molecular view of the contents of the PDB file may not reveal all the details of biological function. The PDB may contain the tertiary structure, but the biologically active quaternary structure is constructed by applying crystallographic symmetry. Viruses are classic examples of this phenomenon. The contents of the asymmetric unit may be one or more capsid proteins, but the complete capsid coat surrounding the RNA is derived by applying crystallographic symmetry to the contents of the asymmetric unit, thereby revealing the biologically active molecule. As expected, resources have been established to deal with quaternary structure. *Protein Quaternary Structure* (*PQS*; Hendrick & Thornton, 1998; http://pqs.ebi.ac.uk/) provides automated quaternary structure derivation and views, and *Virus Particle Explorer* (*VIPER*; http://mmtsb.scripps.edu/viper/viper.html; Reddy *et al.*, 2001) is an example of a virus-specific resource.

For views highlighting features of the molecule the Jena Image Library of Biological Macromolecules is a comprehensive resource (Reichert & Sühnel, 2002; http://www.imb-jena.de/IMAGE.html). Visualization extends beyond traditional ribbon, ball-and-stick and CPK depictions, for example to map physicochemical properties to the sequence and structure, provide surface properties such as electrostatics and show secondary structure, disulfide linkages, salt bridges *etc.* relative to the sequence. Table 3 provides a list of resources having these features as well as a pointer to further information.

2.2.4. Protein families. Often, macromolecular structure forms part of a larger study on a particular family of proteins that are functionally related and leads to a community resource. Individual research laboratories usually develop such resources with interest in specific protein families. The general notion is to be narrow but deep, in contrast to resources like the PDB that are broad but shallow with respect to their information content. Stated another way, the PDB contains a limited amount of information on all macromolecular structures; resources such as the Protein Kinase Resource (PKR; http://pkr.sdsc.edu; Smith *et al.*, 1997) integrate structure as part of additional information on a specific protein family. In the case of PKR, this extends to detailed sequence classifications, motif recognition and relationship to disease, with the overall goal of providing comprehensive information on an important class of enzymes involved in cell signaling. Similar resources to PKR exist for chaperonins, the P450 family, cytokines, esterases, g protein coupled receptors, gluvoamylases, HIV proteases, kinesins, thyroid hormone receptors, topoisomerases and viruses. A more complete list and associated web links can be found at the CMS Molecular Biology Resource (http://restools.sdsc.edu/biotools/biotools25.html).

2.2.5. Genomic relationships. The rate of complete genome sequencing is nothing short of astounding. The impact that this will have on biology through comparative genomics is now being felt (Koonin *et al.*, 2000). At the time of writing, over 800 complete or partial genome sequences can be found in the NCBI genome resource (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = Genome). All three main domains of life – bacteria, archaea and eukaryota – are represented, as well as many viruses and organelles. A natural question that follows with all these genomic data in hand is what is the relationship to protein structure? More specifically, one line of inquiry might relate to protein folds. How abundant is a particular protein fold in a particular genome? Are the distribution of folds equivalent across the major domains of life? These are weighty questions indeed and questions that assume prerequisite knowledge of (i) the complete set of protein folds and (ii) accurate structures of all proteins in a genome. Clearly, we may never know all the folds or all the protein structures from a given genome, but we will come close. Estimates of the number of folds vary from 1000 (Wolf *et al.*, 2000) up to 5000 depending on, among other factors, how the fold is defined. At present, the number of unique folds in the PDB is, again depending on how you classify a fold, approximately 500. Thus, we currently have in hand somewhere between 10 and 50% of all known folds. Efforts are under way to solve all structures in complete genomes; for example, the thermophile *Thermotoga maritimea* (http://www.jcsg.org). Even at this stage, answers to the questions about distribution of structure across complete genomes are asked. Perhaps the best source for this information is Partslist (http://bioinfo.mbb.yale.edu/partslist/) from Mark Gerstein's group at Yale (Qian *et al.*, 2001). Partslist provides details of the distribution of SCOP domains across a number of complete genomes.

2.3. Dynamics

Proteins are not static but dynamic, a feature which is often vital to how they function. Perhaps the clearest examples of these types of motions comes from enzyme–substrate binding, whereupon the enzyme will change its conformation significantly upon substrate binding. While there are many studies on individual proteins, notably those that relate to protein–drug interactions, perhaps the best resource for estimates on the motions in proteins across the complete PDB comes from the morph server (Krebs & Gerstein, 2000; http://bioinfo.mbb.yale.edu/MolMovDB). The morph server, when given a start and end point, will attempt to interpolate

between the two positions to indicate the possible range of motions.

2.4. Protein interactions

Proteins do not function in isolation, but rather in complex networks of protein–protein interactions, protein–DNA/RNA and protein–ligand interactions. While macromolecular structures provide valuable data on such complexes, illustrating how proteins interact in detail, much more data are available than those found in the PDB. Table 4 provides sources of further information containing structural information. Farther removed are the many resources providing complete interaction maps for a complete organism or for complete biochemical pathways; see, for example, http://www.genome.ad.jp/kegg/kegg4.html maintained by the Kyoto Encyclopedia of Genes and Genomes (KEGG).

For details of specific protein–ligand interactions of structures found in the PDB, Relibase (Hendlich, 1998; Bergner *et al.*, 2002; http://relibase.ccdc.cam.ac.uk/) provides features such as ligand searching through text, two-dimensional and three-dimensional substructure searching, similarity searching and automatic superposition of related binding sites to compare ligand-binding modes, water positions, ligand-induced conformational changes *etc.*

2.5. Structure prediction and modeling resources

Existing structures provide a rich source of information for the prediction of protein structure and modeling from primary protein sequence. While the cost of structure determination is decreasing rapidly, it will never keep pace with the cost of sequencing. Hence, the ratio of the number of structures to the number of sequences will remain at several orders of magnitude. Yet, as the number of structures continues to rise, they provide a rich source of template information for structure prediction using techniques such as homology modeling and threading. Such progress is monitored by the Critical Assessment of Structure Prediction (CASP) experiments that are run every two years and in which these methods are compared and hotly debated (http://predictioncenter.llnl.gov/casp4/; Venclovas *et al.*, 1999). Such modeling can be performed in one dimension (secondary structure, solvent accessibility), two dimensions (inter-residue distances) and three dimensions (*ab initio* prediction, homology modeling and threading). Resources even exist to evaluate prediction servers (see, for example, EVA, http://cubic.bioc.columbia.edu/eva/, and LiveBench, http://bioinfo.pl/LiveBench/). To facilitate these prediction efforts, if the depositor permits, sequences of solved protein structures are released ahead of the structures by the PDB to permit unbiased experiments from a continuous source of new targets (see http://www.rcsb.org/pdb/status.html). Other sources of targets are the structural genomics projects which each provide a list of the structures they are working on, including the sequences, which are compiled and updated by the PDB (http://targetdb.pdb.org/). In this case, however, there is no knowing whether a structure will result.

Table 3
Popular resources visualizing macromolecular structures.

Resource	Details
Jena image library	Images depicting biological function and useful links to other resources, http://www.imb-jena.de/IMAGE.html (Reichert & Sühnel, 2002).
PDBSum	Summaries for all protein structures including protein-ligand interactions, http://www.biochem.ucl.ac.uk/bsm/pdbsum/
NDB atlas	Protein–DNA complexes, http://ndbserver.rutgers.edu/NDB/NDBATLAS/
STING	Sequence and property browser, http://mirrors.rcsb.org/SMS/
GRASS	Static *GRASP* images of electrostatic and surface properties http://trantor.bioc.columbia.edu/GRASS/surfserv_enter.cgi
General	World Index of Molecular Visualization resources, http://molvis.sdsc.edu/visres/

Table 4
Popular resources of protein interactions.

Resource	Details
DIP	Database of interacting proteins, http://dip.doe-mbi.ucla.edu/ (Xenarios *et al.*, 2002)
BIND	The biomolecular interaction network database, http://www.bind.ca/ (Bader *et al.*, 2001)
MINT	Molecular interactions database, http://tweety.elm.eu.org/mint/index.html
Relibase	Protein–ligand interactions database, http://relibase.ccdc.cam.ac.uk/ (Hendlich, 1998; Bergner *et al.*, 2002)

Another class of modeling is based on known structures and includes secondary-structure assignment and characterization of topology. See http://restools.sdsc.edu/biotools/biotools9.html for a comprehensive list of structure prediction and modeling resources.

3. The future

Structural genomics is perturbing the structural biology landscape. Structural genomics (Burley *et al.*, 1999) implies high-throughput structure determination for purposes ranging from filling in protein fold space to facilitate comparative modeling to determining as many protein structures from a given genome as possible to furthering our understanding of specific disease states or specific biochemical pathways. While the end product may differ, the process is the same and will result in a large number of structures, estimated at 30 000 by 2005 (Bourne, 1999). Many of these structures will be incomplete, having been discarded in a partially completed state, since they were not deemed useful for the goals of a given project. Others will be complete, but for the first time functionally unclassified. The promise of what could come is given in part by the target-registration database described above. At the time of writing, there are over 14 000 targets in this database. Some of which will be solved and further enrich the large variety of databases of derived information described here. While we are faced with new challenges to judge the quality of the structure information available, we will shortly have richer resources from which to study

structure–function relationships, which will surely further our understanding of biological systems. This is a testament not only to those who produce primary structure data, but to all those who have developed and maintained the resources described herein that have made these advances possible.

We thank Helen Berman, John Westbrook and Christine Zardecki for valuable contributions to this manuscript.

References

Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T. & Hogue, C. W. (2001). *Nucleic Acids Res.* **29**, 242–245.

Bergner, A., Gunther, J., Hendlich, M., Klebe, G. & Verdonk, M. (2002). In the press.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Ravichanran, V., Schneider, B., Thanki, N., Padilla, D., Weissig, H., Westbrook, J. D. & Zardecki, C. (2002). *Acta Cryst.* D**58**, 899–907.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bourne, P. E. (1999). *Bioinformatics*, **15**, 715–716.

Brenner, S. E., Koehl, P. & Levitt, M. (2000). *Nucleic Acids Res.* **28**, 254–256.

Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W. & Swaminathan, S. (1999). *Nature Genet.* **23**, 151–157.

Dietmann, S. & Holm, L. (2001). *Nature Struct. Biol.* **8**, 953–957.

Gibrat, J.-F., Madej, T. & Bryant, S. H. (1996). *Curr. Opin. Struct. Biol.* **6**, 377–385.

Hendlich, M. (1998). *Acta Cryst.* D**54**, 1178–1182.

Hendrick, K. & Thornton, J. M. (1998). *Trends Biochem. Sci.* **23**, 358–361.

Hobohm, U. & Sander, C. (1994). *Protein Sci.* **3**, 522–524.

Holm, L. & Sander, C. (1996). *Science*, **273**, 595–603.

Hooft, R. W., Sander, C. & Vriend, G. (1997). *CABIOS*, **13**, 425–430.

Jones, S. & Thornton, J. M. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 13–20.

Jones, S., van Heyningen, P., Berman, H. M. & Thornton, J. M. (1999). *J. Mol. Biol.* **287**, 877–896.

Kleywegt, G. J. & Jones, T. A. (1995). *Structure*, **15**, 535–540.

Koonin, E. V., Aravind, L. & Kondrashov, A. S. (2000). *Cell*, **101**, 573–576.

Krebs, W. G. & Gerstein, M. (2000). *Nucleic Acids Res.* **28**, 1665–1675.

Laskowski, R. A., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L. & Thornton, J. M. (1997). *Trends Biochem. Sci.* **22**, 488–490.

Lesk, A. M. & Chothia, C. (1980). *J. Mol. Biol.* **136**, 225–270.

Levinthal, C., Barry, C. D., Ward, S. A. & Zwick, M. (1968). *Emerging Concepts in Computer Graphics*, edited by D. Secrest & J. Nievergelt, pp. 231–253. New York: W. A. Benjamin Inc.

Li, W., Jaroszewski, L. & Godzik, A. (2001). *Bioinformatics*, **17**, 282–283.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). *Proteins*, **12**, 345–364.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *Structure*, **5**, 1093–1108.

Qian, J., Stenger, B., Wilson, C. A., Lin, J., Jansen, R., Teichmann, S. A., Park, J., Krebs, W. G., Yu, H., Alexandrov, V., Echols, N. & Gerstein, M. (2001). *Nucleic Acids Res.* **29**, 1750–1764.

Reddy, V. S., Natarajan, P., Okerberg, B., Li, K., Damodaran, K. V., Morton, R. T., Brooks, C. L. III & Johnson, J. E. (2001). *J. Virol.* **75**, 11943–11947.

Reichert, J. & Sühnel, J. (2002). *Nucleic Acids Res.* **30**, 253–254.

Richardson, J. S. (1985). *Methods Enzymol.* **115**, 359–380.

Shindyalov, I. N. & Bourne, P. E. (1998). *Protein Eng.* **11**, 739–747.

Shindyalov, I. N. & Bourne, P. E. (2000). *Proteins*, **38**, 247–260.

Siddiqui, A. S. & Barton, G. J. (1995). *Protein Sci.* **4**, 872–884.

Smith, C., Gribskov, M., Shindyalov, I. N., Taylor, S. S., Ten Eyck, L., Veretnik, S. & Bourne, P. E. (1997). *Trends Biochem. Sci.* **22**, 444–446.

Sowdhamini, R., Burke, D. F., Huang, J.-F., Mizuguchi, K., Nagarajaram, H. A., Srinivasan, N., Steward, R. E. & Blundell, T. L. (1998). *Structure*, **6**, 1087–1094.

Tate, J. (2002). In *Molecular Visualization in Structural Bioinformatics*, edited by E. H. Weissig & P. E. Bourne. New York: Wiley.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.

Venclovas, C., Zemla, A., Fidelis, K. & Moult, J. (1999). *Proteins*, **3**, 231–237.

Weissig, H. & Bourne, P. E. (1999). *Bioinformatics*, **15**, 807–831.

Wolf, Y. I., Grishin, N. V. & Koonin, E. V. (2000). *J. Mol. Biol.* **299**, 897–905.

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. & Eisenberg, D. (2002). *Nucleic Acids Res.* **30**, 303–305.