BRIEFINGS IN FUNCTIONAL GENOMICS AND PROTEOMICS. VOL 6. NO 3. 220–239 doi:10.1093/bfgp/elm020

A comparative view at comprehensive information resources on three-dimensional structures of biological macro-molecules

Rolf Hühne, Frank-Thomas Koch and Jürgen Sühnel

Advance Access publication date 23 October 2007

Abstract

The rapidly increasing amount of information on three-dimensional (3D) structures of biological macro-molecules has still an insufficient impact on genome analysis, functional genomics and proteomics as well as on many other fields in biomedicine including disease-related research. There are, however, attempts to make structural data more easily accessible to the bench biologist. As members of the world-wide Protein Data Bank (wwPDB), the RCSB Protein Data Bank (PDB), the Protein Data Bank Japan and the Macromolecular Structure Database are the primary information resources for 3D structures of proteins, nucleic acids, carbohydrates and complexes thereof. In addition, a number of secondary resources have been set up that also provide information on all currently known structures in a relatively comprehensive manner and not focusing on specific features only. They include PDBsum, the OCA browser-database for protein structure/function, the Molecular Modeling Database and the Jena Library of Biological Macromolecules—JenaLib. Both the primary and secondary resources often merge the information in the PDB files with data from other resources and offer additional analysis tools thereby adding value to the original PDB data. Here, we briefly describe these resources from a user's point of view and from a comparative perspective. It is our aim to guide researchers outside the structure biology field in getting the most out of the 3D structure resources.

Keywords: *three-dimensional structure; proteins; nucleic acids; X-ray crystallography; NMR spectroscopy*

INTRODUCTION

The last 20 years have seen a dramatic increase in the number of known experimental structures of proteins and nucleic acids and other molecules of life. For example, in 1986 the total number of entries in the Protein Data Bank (PDB) was 214 and this number has increased to 44 700 structures as of 17 July 2007 [1,2]. This huge amount of structural information is not yet fully utilized in functional genomics and proteomics. The relatively small impact of structural information outside the structural biology community has been nicely described in a joke by NCBI's Steve Bryant mentioned in an Editorial in *Nature Structural Biology* in 1997 [3]: What do molecular biologists fear most? Three letters: PDB.

The RCSB PDB [1,2] and more recently the world-wide Protein Data Bank (wwPDB) [4] are the single world-wide resources for three-dimensional (3D) structural information on biological macro-molecules. In wwPDB (www.wwpdb.org) organizations that act as deposition, data processing and distribution centres for PDB data have joined forces. The founding members are RCSB PDB (USA), Macromolecular Structure Database (MSD-EBI, Europe) and Protein Data Bank (PDBj, Japan).

Corresponding author. Jürgen Sühnel, Biocomputing Group, Leibniz Institute for Age Research, Fritz Lipmann Institute (FLI), Jena Centre for Bioinformatics, Beutenbergstr 11, D-07745 Jena/Germany. Tel: +49-3641-656200; Fax: +49-3641-656210; E-mail: jsuehnel@fli-leibniz.de

Rolf Hühne is a Postdoc in the FLI Biocomputing Group.
Frank-Thomas Koch is a Postdoc in the FLI Biocomputing Group.
Jürgen Sühnel is Head of the FLI Biocomputing Group.

The Biological Magnetic Resonance Bank (BMRB, USA) [5] group joined the wwPDB in 2006.

The PDB has been established in 1971 and thus belongs to the oldest biological databases [1,2]. Originally, it was a simple file directory with ASCII files in the so-called PDB format. With the advent of the World-Wide Web this situation has dramatically changed and much more user-friendly resources have been established.

The first steps towards more user friendliness were done by simple image collections provided by SWISS-3DIMAGE [6] (www.expasy.ch/sw3d/) and by the IMB Jena Image Library of Biological Macromolecules [7] (www.imb-jena.de/IMAGE.html) as well as by browsers provided at that time by the PDB [6,8,9]. SWISS-3DIMAGE has not been further developed but the images are still available via the UniProt database [10]. The PDB has undertaken significant efforts to improve the user-friendliness of data access [2] and also the IMB Jena Image Library has been completely redesigned and extended since 1993. In 2005, due to a name change of the hosting institute it has changed its name to Jena Library of Biological Macromolecules—JenaLib (www.fli-leibniz.de/IMAGE.html). Later, PDBsum (www.ebi.ac.uk/pdbsum/) [11], OCA (ispc.weizmann.ac.il/oca-docs/oca-home.html), the PDBj (www.pdbj.org) [12] and the MSD (www.ebi.ac.uk/msd/) [13] were set up.

In addition to the comprehensive resources mentioned, there are other 3D structure databases that either offer information on a subset of the currently known structures of biological macromolecules, such as the Nucleic Acid Database (NDB) [14], RNABase [15], the BMRB [5] and the Electron Density Server—EDS [16] or provide information on relatively specific features such as the PDBREPORT database [17] or the STING server [18]. The description of these latter databases is beyond the scope of this article.

There is a substantial overlap between the data and services offered by the comprehensive resources. However, many of them do also provide unique information that is not available in the others. Information on the services as well on the specific strengths and weak points is not easy to get for a potential user. The features of most of the comprehensive resources have been reported either in the database issue of *Nucleic Acids Research* or elsewhere. Thus far, there is, however, only one attempt of a comparative description dating back to 2002 [19]. Recently, some of the 3D structure databases have been redesigned and significantly extended. Therefore, we think it is a good time for an updated review adopting a comparative perspective and focusing on information that is unique to a particular resource.

COMPREHENSIVE 3D STRUCTURE DATABASES

The description is started with the wwPDB databases followed by the other resources. The databases are listed in reverse alphabetical order. For each of the databases there is a figure showing either the full or a part of the atlas page. This provides an at-a-glance view of many features. For a few resources additional figures are displayed. In addition, in Table 1, database characteristics are listed and compared. Table 2 offers information on a subset of these data focusing on unique features. Recent developments in the JenaLib database were not yet described elsewhere. Therefore, its features are described in more detail as compared to the other resources. Finally, it should be noted that the description of individual databases cannot be comprehensive. We are here primarily concerned with atlas pages offered by all resources but also mention occasionally other analysis tools.

The RCSB Protein Data Bank (PDB)

The main PDB archive consists of structures determined using experimental methods only. Models, accepted prior to 15 October 2006, are stored in a separate archive. The RCSB PDB offers structure information on five different pages titled Structure Summary, Biology & Chemistry, Materials & Methods, Sequence Details and Geometry (Figure 1). In addition to the PDB data, the Structure Summary page provides also SCOP [20], CATH [21] and PFAM [22] information. Given, there is a difference between asymmetric and biological units the user can switch between both representations. For nucleic acid containing structures the thumbnail image is taken from the NDB. Interactive visualization is possible via KiNG, Jmol (www.jmol.org), WebMol [23], the MBT Protein Workshop, the MBT (Molecular Biology Toolkit) Ligand Explorer, QuickPDB, RasMol [24] and the Swiss-PDB Viewer [25]. There is also an option for creating high-resolution images. The Chemistry & Biology page offers detailed information on the molecule under study. This includes

Table I: Features available in the PDB, PDBj, MSD, PDBsum, OCA, MMDB and JenaLib databases

		PDB	PDBj	MSD[a]	PDBsum	OCA	MMDB	JenaLib
ID Search	PDB ID	+	+	+	+	+	+	+
	NDB ID (PDR001: 3cro)	+	+	−	−	−	−	+
	PubMed ID (6941276, Ibna)	+	−	+	−	+	−	−
	UniProt ID (LYSC.CHICK: 5lyz, . . .)	+	−	+	−	−	−	+
	(P00698: 5lyz, . . .)							
		+	−	+	−	−	−	+
	EC ID (3.6.1.1: Inorganic diphosphatase)	+	+	+	+	+	+	+
	GO ID (GO:0003823: antigen binding) . . .)	−	−	+	−	+	−	+
	CATH ID (10mhA01: 10mh)	−	−	−	−	+	−	−
	(3.40.50: Rossmann fold)							
		−	−	+ [b]	−	−	−	+
	SCOP ID (19926: 1a2y)	−	−	+	−	−	−	+
	Pfam ID (SH2: SH2 domain)	−	−	−	−	−	−	+
	(PF00017: SH2 domain)							
		−	−	+	−	+	−	+
	InterPro ID (IPR008162: Pyrophosphatase)	−	−	+	−	+	−	−
	SMART ID (FN3: FN3 domain)	−	−	−	−	−	−	+
	(SM00060: FN3 domain)							
		−	−	−	−	−	−	+
	PROSITE ID (PS00059:[c]) (ADH.ZINC:[c])	−	+	−	−	−	−	+
		−	−	−	−	−	−	+
Text search	Single input field (QuickSearch)	+	+	+	+	+	+	+
	Multiple input field form	+	+	+	+	+	−	+
	Complex query builder	+	−	+	−	−	−	−
Sequence search	FASTA	+	−	−/+	+	+	−	−
	BLAST	+	+	−	−	−	+	−
	Structural genomics targets	+	−	+	−	−	−	−
Structure search	Structural similarity	−	+	+	−	−	+	−
Search result display	Highlighted search term hits	−	−	−	+	−	−	+
	Structure image thumbnail	+	−	−	−	−	+	−
	Listed by PDB ID	−	+	+	+	+	−	+
	Listed by rank	+	−	−	−	+	+	−
	Page breaks (max. 50–500 hits per page)	+	+	+	+	+	+	−
Search result processing	Generate PDB ID list	+	−	−/−	−	−	−	+
	Generate customized entry list	+	−	−/+	−	−	−	+
	Generate SCOP/CATH classification tree	+	−	−/−	−	−	−	+
	Change sort order	+	+	+/+	−	−	−	−
	Select results for processing	+	−	−/+	−	+	+	−
	Remove similar structures (sequence identity)	+	−	−/−	−	−	−	−
	View only all structure image thumbnails	+	−	−/−	−	−	−	−
	Download all/selected entries	+	−	+/+	−	+	+	−
Available entries	Experimental entries	+	+	+	+	+	+	+
	Theoretical entries, atlas page	−	−	+	+	+	−	+
	Theoretical entries, PDB file only	+	+	−	−	−	−	−
	Superseded entries, own atlas page	−	−	−	+	−	−	−
	Superseded entries, forward mechanism	+	+	−	+	+	−	+
Atlas page	Hetero component table	+	+	+	+	+	+ (e)	+
	Hetero component 2D structure	+	−	+	+	−	+	−
	Hetero component interaction table	−	−	−	−	+	−	+ (e)
	Hetero component interaction graphics	−	−	+	+	−	−	−
	Active site table	−	+	+	+	−	−	+
	ProMotif motif table	−	−	−	+	−	−	−
	Graphical ProMotif motif view (2D)	−	−	−	+	−	−	−
	Protein–protein interaction table	+ (e)	−	−	+	+	−	+ (e)
	Protein–protein interaction graphics	−	−	−	+	−	−	−
	Membrane orientation	−	−	−	−	+	−	+ (e)
	Protein surface	+ (e)	−	+	+	+ (e)	−	+ (e)
	Cleft table	−	−	−	+	−	−	−
	Structural neighbours	+ (e)	+ (e)	+	−	−	+	+ (e)
	SCOP classification	+	+ (e)	+ (e)	+	+	−	+

(continued)

Table I: Continued

		PDB	PDBj	MSD[a]	PDBsum	OCA	MMDB	JenaLib
	CATH classification	+	+ (e)	+ (e)	+	+ (e)	−	+
	PFAM classification	+	−	+ (e)	+	+ (e)	+	+ (e)
	GeneOntology annotation	+	+	+ (e)	+	+	−	+
	SAP(SNP)/variant mapping	+	−	−	−	−	−	+
	PROSITE motif mapping	−	+	+	+	−	−	+
	PDB sequence view	+	+	+	+	+	+	+
	UniProt sequence alignment view	−	−	+	−	−	−	+
	PFAM sequence alignment view	−	−	−	−	+ (e)	+	−
	Related entries	+	+	+	−	+	+	+
	Genetic source/genome information	+	−	−	−	+	−	+ (e)
	Crystallographic data	+	+	+	−	+	−	−
	Structure validation	+ (e)	−	−	+	+ (e)	−	+ (e)
	Enzyme information	+ (e)	+	+ (e)	+	+	−	+ (e)
	Disease information	+	−	−	−	+	−	+ (e)
	Age related information	−	−	−	−	−	−	+ (e)
	DNA bending	−	−	−	−	−	−	+
	Protein disorder	−	−	−	−	−	−	+ (e)
	Geometry statistics (e.g.: bond lengths/angles)	+	−	−	−	−	−	−
	Number of external crosslinks—example: Ideh	44	13	15	30	30	7	50
3D Viewer	AstexViewer (Java applet)	−	−	+	+ (e)	+	−	+ (e)
	Chime (browser plugin, requires local installation)	−	−	−	−	−	−	+ (e)
	Cn3D (programme, requires local installation)	−	−	−	−	−	+	−
	First Glance (Jmol−based Java applet)	−	−	−	−	+	−	−
	JenaLib Jmol Viewer (Java applet, Javascript)	−	−	−	−	−	−	+
	Jmol (Java applet)	+	−	−	+	+	−	+
	jV (Java applet, requires JOGL)	−	+	−	−	−	−	−
	KiNG	+	−	−	−	−	−	−
	MBT Ligand Explorer (Java program)	+	−	−	−	−	−	−
	MBT Protein Workshop (Java program)	+	−	−	−	−	−	−
	MBT Simple Viewer (Java program)	+	−	−	−	−	−	−
	QuickPDB (Java applet)	+	−	−	−	−	−	−
	RasMol (programme, requires local installation)	+	−	+	+	+	+	+
	Swiss−PDB Viewer (programme, requires local installation)	+	−	−	−	−	−	−
	WebMol (Java applet)	+	−	−	−	−	−	+ (e)
3D Visualization options	Overall structure	+	+	+	+	+	+	+
	Hetero components	+	+	+	+	+	+	+
	Active sites	+	−	+	−	−	−	+
	SAPs(SNPs)/variants	+	−	−	−	−	−	+
	PROSITE motifs	−	−	−	−	−	−	+
	SCOP domains	−	−	−	−	+	−	+
	CATH domains	−	−	−	−	−	−	+
	Dipole moments	−	−	−	−	+	−	−
	Non-covalent bond finder	−	−	−	−	+	−	−
	Asymmetric/Biological unit	+	−	+	+	−	−	+
	Biological unit information source	PDB	−	PQS	PQS	PDB	−	PDB
Sequence/Alignment view	PDB sequence	+	−	+	+	+	+	+
	Aligned UniProt sequence	+	−	+	−	−	−	+
	Secondary structure elements	+	−	+	+	−	−	+
	Modified residues	+	−	−	−	−	−	+
	Disulfide bonds	+	−	−	+	−	−	−
	SCOP domains	+	−	−	−	−	−	−
	CATH domains	+	−	−	−	−	−	−
	PROSITE motifs	−	−	−	+	−	−	−
	Active sites	−	−	−	+	−	−	−
	Residue contacts	−	−	−	+	−	−	−
Other features	SCOP classification tree browser	+	−	−	−	−	−	+
	CATH classification tree browser	+	−	−	−	−	−	+
	Gene Ontology tree browser	+	−	−	−	−	−	+
	Genome location tree browser	+	−	−	−	−	−	−

(continued)

Table I: Continued

	PDB	PDBj	MSD[a]	PDBsum	OCA	MMDB	JenaLib
Taxonomy tree browser	+	−	−	−	−	−	−
Hetero component browser	−	−	−	+	−	−	+
PROSITE motif browser	−	−	−	+	−	−	−
Enzyme classification browser	+	−	−	+	−	−	−
MeSH (Medical Subject Headings) browser	+	−	−	−	−	−	−
Customized entry lists	+	−	+	−	−	−	+
PDB file viewer with content highlighting	−	−	−	−	−	−	+

[a]MSDlite and MSDpro are occasionally mentioned separately (MSDlite/MSDpro).
[b]Search for the CATH code (e.g.: 3.10.130.10, 3.40.50) should be possible in MSD but did not work in our case.
[c]Zinc-containing alcohol dehydrogenases signature.
Note that not all features of each individual database are included. External links are indicated by (e). In the ID search section, ID is understood as a unique database identifier that may have a different name in some cases, for example, EC number or UniProt accession number and entry name. The 'IDs' used for searching are indicated in parentheses. Some resources perform a full-text search in the PDB file header and can thus identify IDs contained in the PDB file. A plus is only indicated in the Table, if the corresponding ID is not included in the PDB file (exceptions: PDB ID, EC number).

Gene Ontology [26] and PubMed MESH terms as well as information on associated pathways and catalytic sites. On the Materials & Methods page specific details of the structure determination method, in particular, X-ray or NMR data, are displayed. The Sequence Details page offers an informative view of the sequence including information on secondary structure, disulphide bonds and SCOP domains (Figure 2). Finally, the Geometry page displays data on bond lengths, bond angles and dihedral angles. The atlas page offers also a number of external links and enables the user to display or download PDB files in ASCII, XML and mmCIF format. Navigation through the pages requires reloading.

A unique feature of the RCSB PDB is the Structural Genomics Information Portal with information on worldwide Structural Genomics initiatives, the TargetDB and PepcDB databases that offer information on the progress of structure determination and also a list of functional annotation sites.

The advanced search option enables the building of really complex queries based on the evaluation of sub-queries (Figure 3). It is an extremely powerful search tool.

Last but not least, educational information and site tutorials are also available. This includes, for example, the very well designed Molecule-of-the-Month series.

The Protein Data Bank Japan (PDBj)

The PDBj was set up in 2000 [12]. For experimental structures it provides information on atlas pages, whereas for theoretical models only links to the PDB files are available. This is in line with the current PDB policy. There is a summary page and in addition pages on structural details, experimental details, function details, sequence neighbour and finally download/display page and links pages (Figure 4). To a large extent the information shown on these pages is taken from the PDB file. However, there are also additional data. The functional details page offers also functional information from Gene Ontology data [26] and from UniProt [10] as well as from PROSITE [27]. Moreover, catalytic information from the databases Catalytic Site Atlas (CSA) [28] and Catalytic Residue Dataset (CATRES) [29] is offered. The sequence neighbour page lists all PDB entries with the same UniProt sequence (exact matches) and also other structures with related sequences. The download/display page contains links for downloading the PDB files in ASCII, XML and mmCIF format as well as structure factors if available. On the summary page, there are mono representations of the structure in three different orientations and two different sizes available. Also, there are visualization options for the electron density map and for the structure with two different Java viewers. They need the Java(TM)Plug-in 1.4 and Java3D 1.3 or JOGL library 1.0.

PDBj has a QuickSearch and more sophisticated search options both based on PDB XML files. In QuickSearch, a search is possible for the PDB ID and for any other text string that obviously not only scans the KEYWD record but performs the search in the complete header of the PDB file. In addition to a more or less standard Advanced Search, there are also more sophisticated tools called XQuery/XPath Search.

Table 2: Selected features of 3D structure databases of biological macromolecules with an emphasis on uniqueness

RCSB PDB
- Geometric analysis of individual structures
- Target lists for structural genomics projects
- Information on to be released entries
- Crystallographic data
- Links to 44 external databases (e.g.: Ideh)
- Biological unit information source: PDB

PDBj
- Crystallographic data (visualization of the electron density map)
- Links to 13 external databases (e.g.: Ideh)
- Biological unit information source: PDB

MSD
- Analysis tools: MSDlite, MSDpro, MSDmotif, MSDtemplate, MSDpisa, MSDchem, MSDmine, MSDsite, MSDfold, MSDtarget and MSDanalysis
- Structural genomics targets
- Structural comparison
- UniProt/PDB alignment
- Links to 15 external databases (e.g.: Ideh)
- Biological unit information source: PQS

PDBsum
- Upload option for own structure files for image generation
- Author-generated images
- Cleft table
- LIGPLOT analysis for ligand–protein interaction
- NUCPLOT analysis for protein–nucleic acid interaction
- Links to 30 external databases (e.g.: Ideh)
- Biological unit information source: PQS

OCA
- Extensive disease-related information
- Automatic access, returning XML and plain text results that allow easy integration into other software, web servers or batch queries from large sequencing or proteomics centers
- Ligand-protein contact analysis
- Analysis of contacts of structural units
- Crystal contacts
- Links to 30 external databases (e.g.: Ideh)
- Biological unit information source: PDB

MMDB
- Structural comparison
- Links to 7 external databases (e.g.: Ideh)
- Biological unit information source: None

JenaLib
- Integrated visualization of SCOP and CATH domains, single amino acid polymorphisms and PROSITE motifs (work on exon structure in progress)
- Fully customizable entry list generation
- Generation of mono and stereo MolScript vector graphics images (PDF)
- Age-related information
- UniProt/PDB alignment
- Links to 50 external databases (e.g.: Ideh)
- Biological unit information source: PDB

The Macromolecular Structure Database (MSD)

The MSD at the EBI offers information on experimental and theoretical structures and also provides atlas pages for superseded entries [13]. The information for a particular entry is displayed on five different pages titled Summary, Assembly, Sequence, Citation, Similarity and Visualization (Figure 5). The advantage of this data organization is that the information is clearly arranged. On the other hand, it is a disadvantage that the server has to be contacted to (re)load a page when navigating through the pages, thereby slowing down the work. Assembly information is not taken from the PDB as in the other wwPDB resources but from the Protein Quaternary Structure Server PQS [30].

The sequence page offers a UniProt/PDB alignment (Figure 6). The citation page provides a

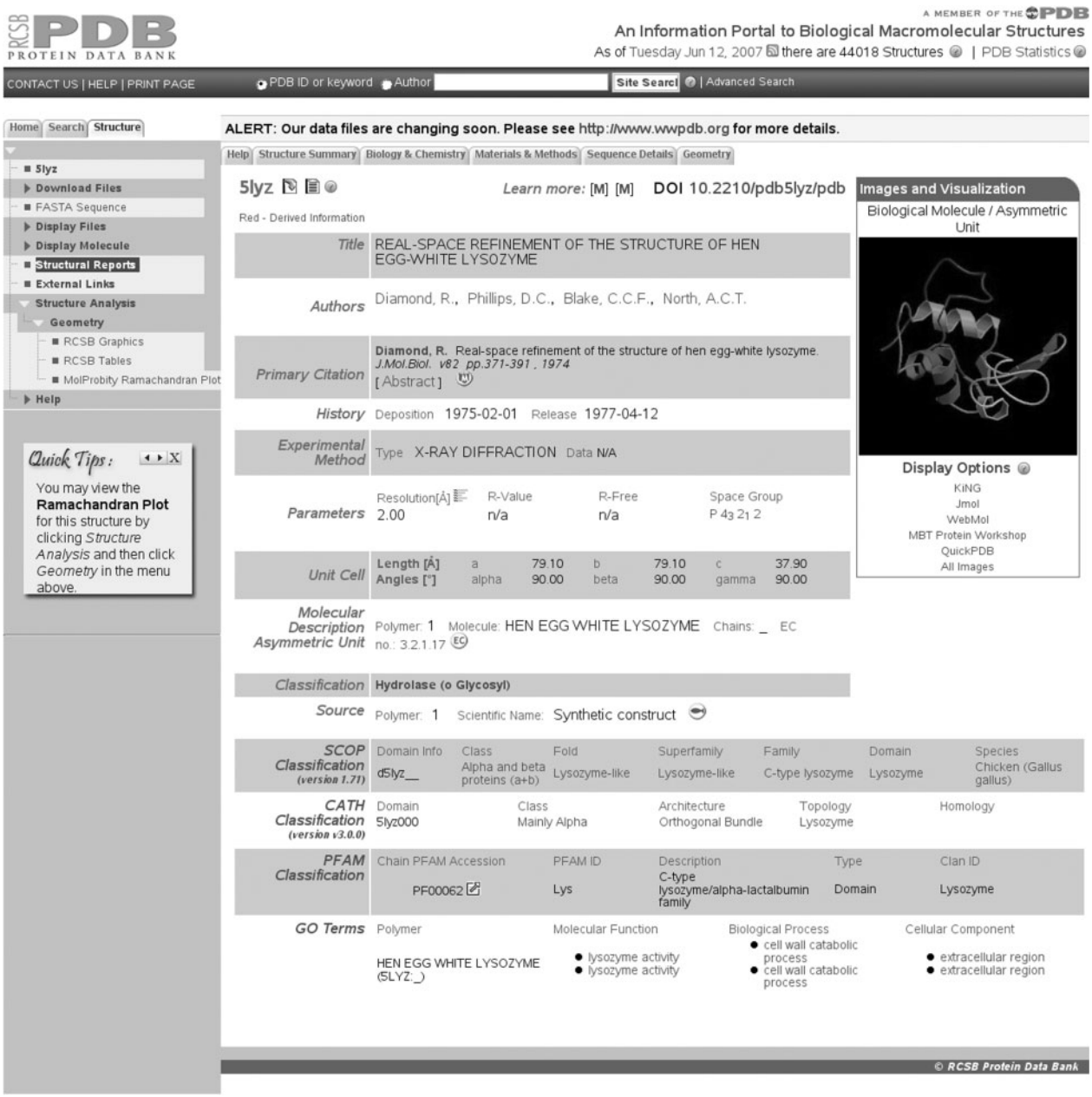

Figure 1: RCSB PDB atlas page for lysozyme (5lyz).

complete list of primary and secondary citations of a particular entry as well as of references for related entries. On the similarity page, an especially useful feature is the structural comparison to all other PDB structures via the MSD fold system. Finally, the visualization page offers the Astex-Viewer, RasMol and a JenaLib link as visualization options. MSD offers an impressive number of additional analysis tools called MSDlite, MSDpro, MSDmotif, MSDtemplate, MSDpisa, MSDchem, MSDmine, MSDsite, MSDfold, MSDtarget and MSDanalysis. To mention only a few: both MSDlite and MSDpro represent excellent search interfaces, MSDsite is a database search and retrieval system for the analysis of ligands and active sites [31], MSDmotif tries to identify small motifs ocurring in 3D structures and MSDfold is a structure matching tool.

PDBsum

The PDBsum database was created at the University College of London in 1995 and has now moved to the European Bioinformatics Institute [32]. Database entries can be accessed either by the PDB ID or a text search in the TITLE, HEADER, COMPND,

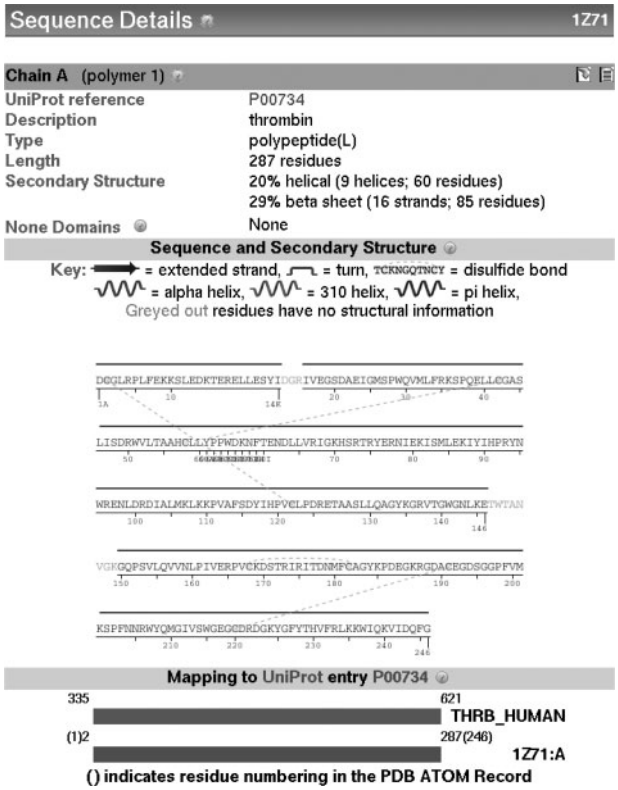

Figure 2: RCSB PDB Sequence details page for a thrombin/hirudin IIIB complex (1z71). Only chain A is shown.

SOURCE and AUTHOR records of the PDB files or also by sequence.

The database can also be browsed by a number of lists. They include PDB codes, ligands, the Enzyme Classification (E.C.) scheme, PROSITE patterns, species and a highlights page with various categories of entries such as the oldest, newest, largest and smallest ones, for example. Directly from the atlas page a PROCHECK job can be started [33].

On top of the atlas page, both asymmetric and biological unit thumbnails are shown for crystallographic structures (Figure 7). Interactive visualization is possible via Jmol (www.jmol.org), RasMol [24] and the Astex Viewer. There is also an option with RasMol for structure orientation and Molscript [34] and Raster3D [35] for the generation of high-quality images. Finally, own structures can be uploaded and visualized with the PDBsum tools. This latter option is not available in other resources.

Compact information on chains, ligands and solvent molecules is given on the atlas page. PDBsum offers a very nice view of sequences with information on secondary structure, residue conservation and residue contacts to the nucleic acid part in the case of protein–nucleic acid complexes (Figure 8). There is also a topology view for proteins and a NUCPLOT [36] view for nucleic acid parts

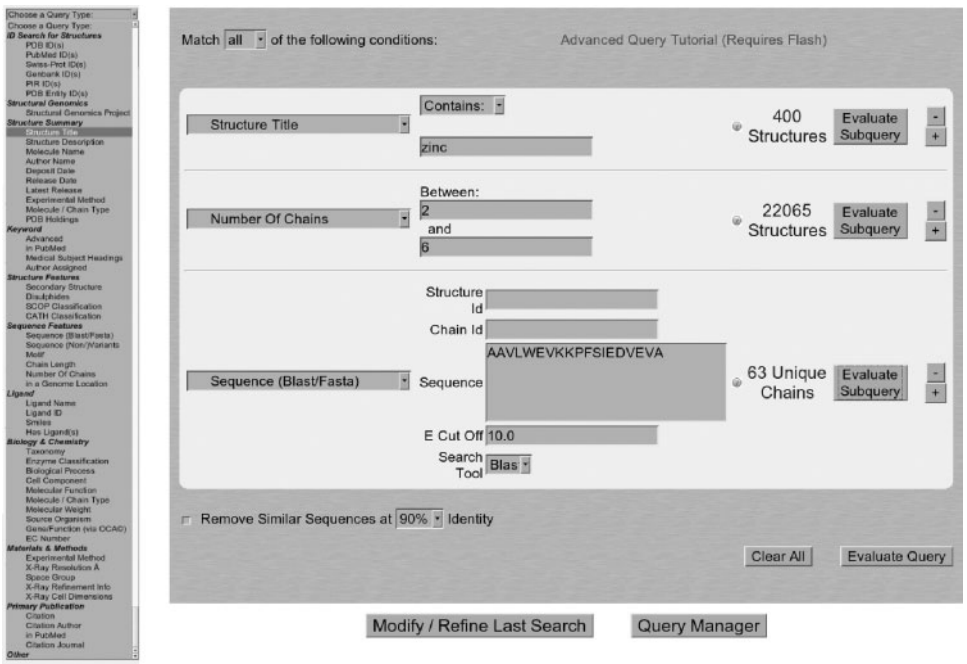

Figure 3: RCSB PDB advanced search option. The example shows the combination of three subqueries. On the left all query types are shown.

Figure 4: PDBj atlas page for lysozyme (5 lyz).

interacting with a protein. Finally, surface representation is available and a cleft analysis is also provided.

A very interesting and unique PDBsum feature is the availability of key figures extracted from the literature [37]. In many cases, author generated images have much higher information content than other ones. So, this is really an extremely useful and informative feature.

The OCA database

The OCA database for protein structure and function is hosted by the Weizmann Institute of Science. OCA offers both a PDBlite search and an advanced search option. In the advanced search gene and disease names can, for example, be used and there is also a possibility for a sequence search. An advantage of the search functionality of OCA is the usage of a thesaurus and an automatic spelling verification in the query mechanism. Database entries can be accessed either by the PDB ID or a text search in the TITLE, HEADER, COMPND, SOURCE and AUTHOR records of the PDB files or also by sequence.

The structure information for an entry is contained in a single page (Figure 9). There is a detailed compound description. The remaining part is divided into a number of subsections titled Data retrieval, View in 3D, Visual 3D analysis, Structure-derived information, Sequence-derived information and Movements, Other resources and Movements, Movies and Images. 3D structure visualization is possible via Jmol (www.jmol.org), AstexViewer and RasMol [24]. Both CATH [21] and SCOP [20] information is part of the atlas page. Information on SCOP domains can also be visualized in the 3D structure by Jmol and RasMol. A particular OCA strength is the relatively detailed disease information.

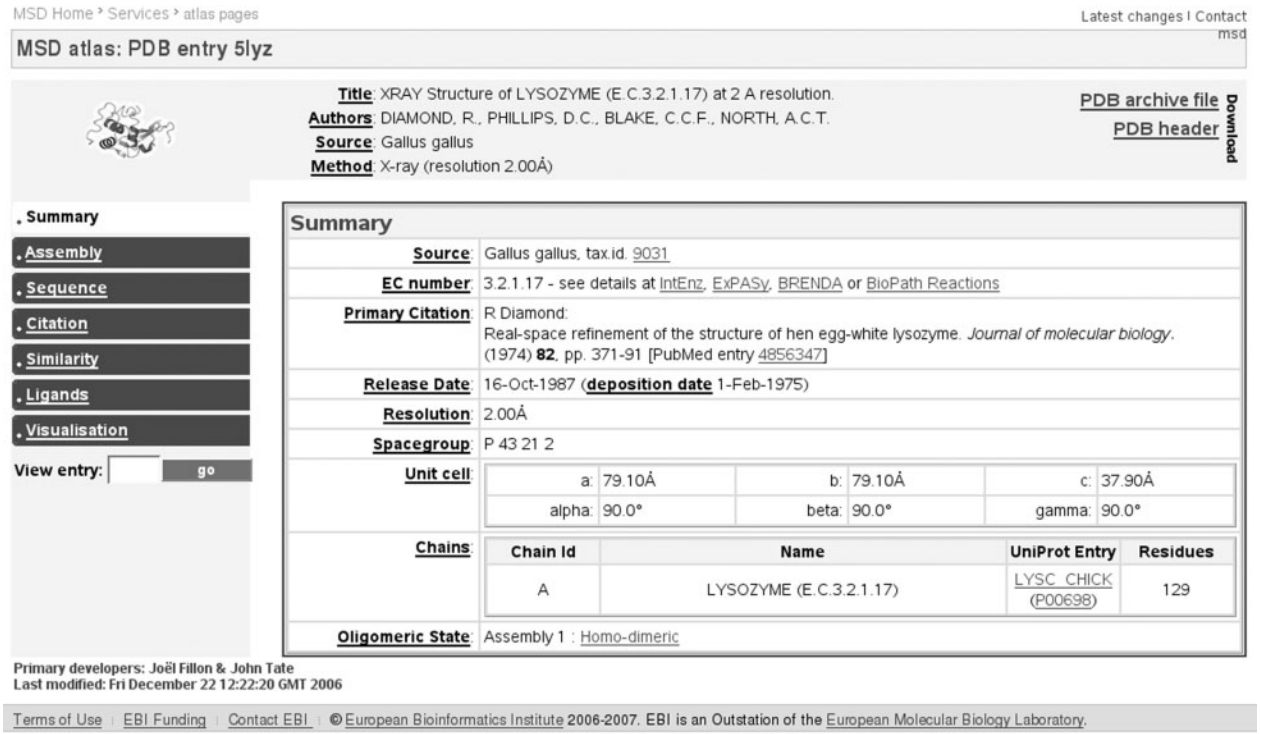

Figure 5: MSD atlas page for lysozyme (5 lyz).

Figure 6: MSD UniProt/PDB alignment view for a thrombin/hirudin IIIB complex (1z71).

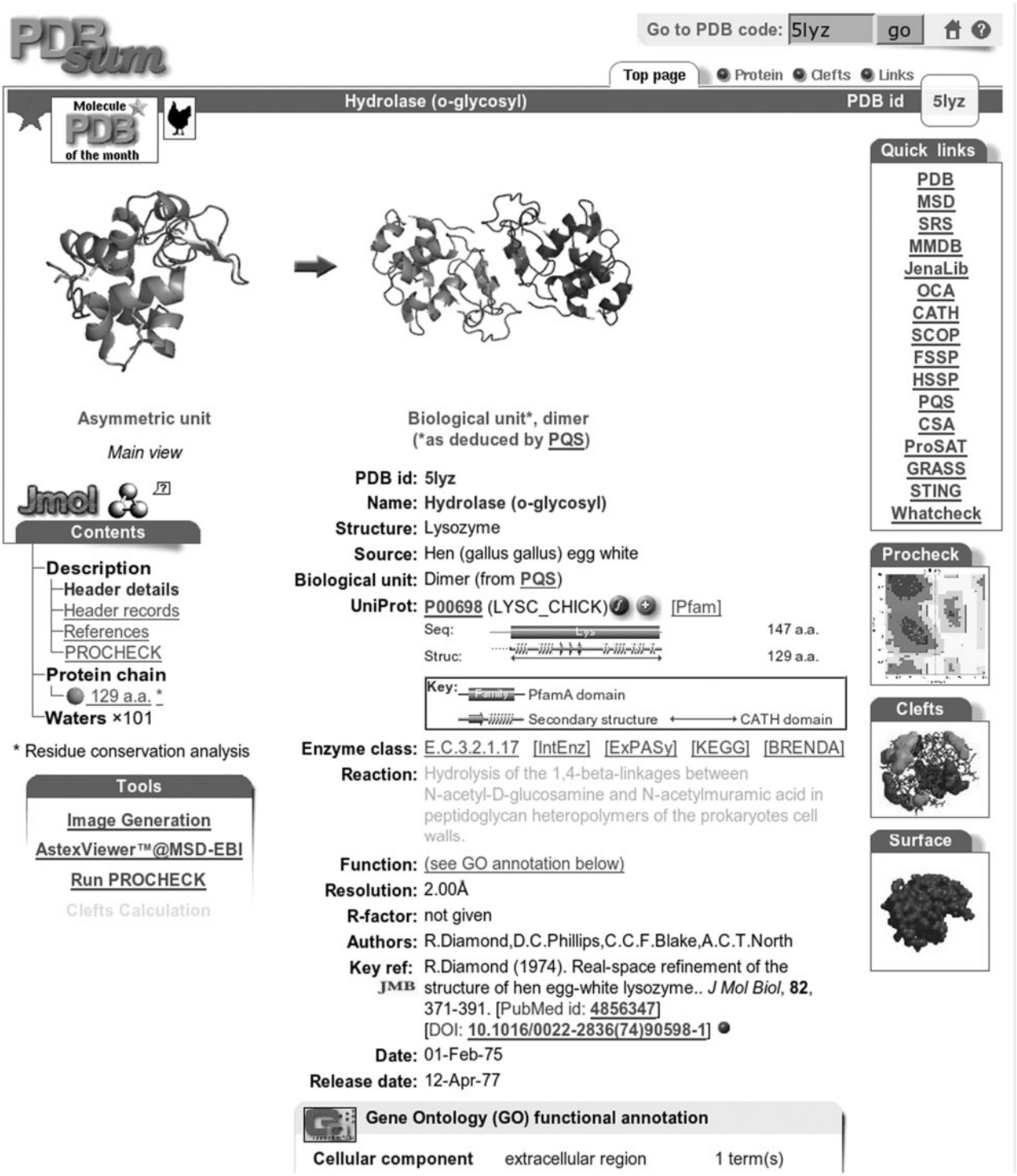

Figure 7: PDBsum atlas page for lysozyme (5 lyz).

OCA is open to be used for automatic access, returning XML and plain text results that allow easy integration into other software, web servers or batch queries from large sequencing or proteomics centres. A list of reporting formats and tags is available from bip.weizmann.ac.il/oca-docs/faq.html.

The Molecular Modeling Database (MMDB)

NCBI's structure database is called MMDB and it offers information on all experimentally determined 3D structures of biological macro-molecules [38]. Visualization is possible via the Cn3D viewer

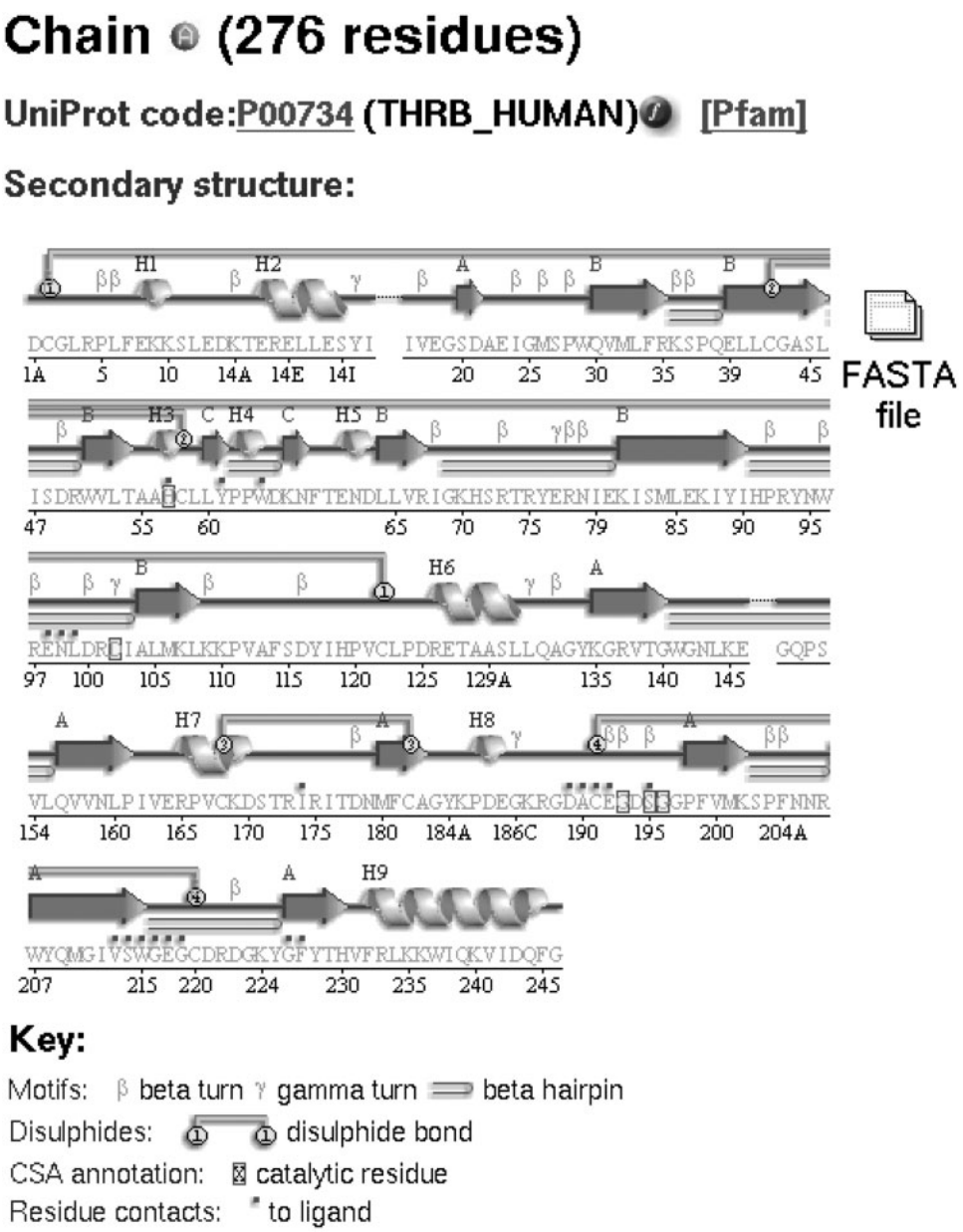

Figure 8: PDBsum sequence and secondary structure information for a thrombin/hirudin IIIB complex (1z71). Only chain A is shown.

(www.ncbi.nlm.nih.gov.ilsprod.lib.neu.edu/Structure/CN3D/cn3d.shtml) and RasMol [24]. As compared to the other databases, the information content of the atlas pages is rather small (Figure 10). A very useful feature is, however, information on structural neighbours derived from the VAST algorithm [39]. Also, due to the usage of the PDBeast system (130.14.29.110/Structure/PDBEAST/pdbeast.shtml) species information is especially reliable and avoids the inconsistencies of the original PDB data.

The Jena Library of Biological Macromolecules (JenaLib)

The JenaLib database was set up in 1993, originally as a gopher-accessible image archive under the name IMB Jena Image Library of Biological Macromolecules [40]. In 1998, a data pipeline was established that combined automatic and manual processing. It generated HTML atlas pages for all database entries. The database consists of two major sections, the atlas of macro-molecule structures and basic information

5LYZ	Hydrolase (O-Glycosyl)	date	Feb 01, 1975
authors	R.Diamond, D.C.Phillips, C.C.F.Blake, A.C.T.North		
compound		source	
Lysozyme (E.C.3.2.1.17)		Hen (Gallus Gallus) Egg White	
symmetry	Space Group: P 43 21 2	R_factor	

crystal cell	length a	length b	length c	angle alpha	angle beta	angle gamma
	79.100	79.100	37.900	90.00	90.00	90.00

method	X-Ray Diffraction	resolution	2.0 Å
ligand		enzyme	Lysozyme;. Muramidase;. Globulin G;. Mucopeptide glucohydrolase;. Globulin G1;. N,O-diacetylmuramidase;. Lysozyme g;. L-7001;. 1,4-N-acetylmuramidase;. Mucopeptide N-acetylmuramoylhydrolase;. PR1-lysozyme. Hydrolase E.C.3.2.1.17 ***BRENDA***
similarity	Belongs to the glycosyl hydrolase 22 family.[Lys]		
subunit	Monomer.		
catalytic activ.	Hydrolysis of 1,4-beta-linkages between n- acetylmuramic acid and n-acetyl-d-glucosamine residues in a peptidoglycan and between n-acetyl-d-glucosamine residues in chitodextrins.		
tissue	In the egg white and polymorphonuclear leukocytes.		
genes	chiA1 (*B. circulans*); CPL1 (*B. Cp-1*); 17, lyz (*B. P1*); 27, 5, E (*B. T4*); LYZ (*G. gallus*); CTS1 (*C. immitis*); LYZ (*C. virginianus*); LYZ (*C. coturnix*); ivy (*E. coli*); CHIT1, LYZ (*H. sapiens*); R (*B. lambda*); LYZ (*M. gallopavo*); Lyz (*M. musculus*); LYZ (*N. meleagris*); LYZ (*P. colchicus*); ivy (*P. aeruginosa*); chiA, chiB (*S. marcescens*); acm (*S. globisporus*)		
function	Lysozymes have primarily a bacteriolytic function; those in tissues and body fluids are associated with the monocyte- macrophage system and enhance the activity of immunoagents.		

Gene Ontology	Chain	Function	Process	Component
	A	• lysozyme activity • catalytic activity • hydrolase activity • hydrolase activity, acting o...	• carbohydrate metabolic proce... • type I hypersensitivity • cell wall catabolic process • cytolysis • defense response to bacteriu...	• extracellular region

disease	Amyloidosis,renal
Primary reference	Real-space refinement of the structure of hen egg-white lysozyme., Diamond R, J Mol Biol 1974 Jan 25;82(3):371-91. PMID:4856347

Data retrieval

- Asymmetric unit, PDB entry: [header only] [complete with coordinates] (36 Kb) [Save to disk]
- Biological Unit Coordinates (5lyz.pdb1.gz) 26 Kb
- CSU: Contacts of Structural Units for 5LYZ
- Likely Quarternary Molecular Structure file(s) for 5LYZ
- Retrieve 5LYZ in mmCIF format [Save to disk]
- SEQRES to COORDINATES correlation for **5LYZ** from S2C, [Save to disk]

View 5LYZ in 3D

- On Jmol, a nice Rasmol like molecule viewer. This is good for easiest viewing of basic structure.
- On FirstGlance, an excellent tool for a guided tour on the structure components, by E. Martz.
- On AstexViewer, from MSD-EBI (viewer documentation).
- On RasMol (Install RasMol freeware) Here's help on how to use RasMol.

Visual 3D analysis of 5LYZ

- Protein Explorer, Easier to use and more powerful than RasMol. (Win and Mac only)
- Noncovalent Bond Finder displays the closest contacts to ligand or interface, reports distances, walks over water bridges.
- STING, shows the **sequence** and maps the locations of sequence residues or residue ranges onto the 3D structure.
- GRASS (Graphical Representation and Analysis of Structure Server)
- Cartoon representation from PDB Cartoon
- Ramachandran plot from PDBSum

Structure-derived information

- Dipole moment, from Dipole Server at Weizmann Institute
- Crystal Contacts, from CryCo at Weizmann Institute
- 3D motif for 5LYZ, from MSDmotif at EBI
- Genome occurence of 5LYZ's fold from GeneCensus
- Classification of representative domains in **scop** (Structural Classification of Proteins)
  - Domain d5lyz__, region [Jmol] [rasmolscript] [script source]
- Fold representative 5lyz from FSSP and Dali (Families of Structurally Similar Proteins)
- Class (fold), Architecture (subfold), Topology, Homologous superfamily from CATH
- Summaries and structural analyses of PDB data files from PDBSum
- Identification of Protein Pockets & Cavities at CASTp

Sequence-derived information

- View one-letter amino acid or nucleotide sequence for each chain: [5lyz_]
- SWISS-PROT database: [Q90884] [P00698]
- **Domain** organization of [LYSC_CHICK] by SWISSPFAM
- **Domain** found in 5LYZ: [LYZ1] by SMART
- **Conserved** protein region description, domain view and alignment from family Lys (PF00062) of Pfam.
- **Alignments** of the sequence of 5LYZ with the sequences similar proteins can be viewed for classification [LYSC_CHICK] at ProtoMap. *Click on "Neighbors List", then on the "See Alignments" button below the list.*
- A sequence distance **tree** ("phylogenetic tree") can be viewed for 1ACL's classification [LYSC_CHICK] at ProtoMap. *Click on the Cluster number.*

Other resources with information on 5LYZ

- InterPro: IPR000974 , IPR001916
- PDBREPORT (protein verification by WHAT_CHECK procedures)
- MMDB (Entrez's Structure Database)

Movements, Movies and Images

- Domain movements in 5LYZ from the Database of Macromolecular Movements.
- Images from IMB Jena Image Library of Biological Macromolecules.

You may enter another PDB ID code Select this ID

Go [Back], to the [PDB Lite page], to the [OCA Search page] or to the [PDB Home page]

OCA© by Jaime Prilusky, 1996-2006
Bioinformatics and Biological Computing
Weizmann Institute of Science

Figure 9: OCA atlas page for lysozyme (5 lyz).

Figure 10: MMDB atlas page for lysozyme (5 lyz).

on the architecture of biopolymers. The latter includes, for example, subsections on experimental methods for structure determination and on nucleic acid nomenclature and structure as well as an amino acid repository.

The JenaLib database provides atlas pages for all entries from the PDB and NDB databases. So, the small number of structures available from the NDB only can also be accessed via the JenaLib database. Both experimental and theoretical structures are included. When searching for superseded entries the user is automatically forwarded to the new structure.

In an attempt of improved data integration a database has been built that contains additional data besides the data from the original PDB files, e.g.: Gene Ontology information [26], structural data from the SCOP [20] and CATH [21] classification schemes, sequence data from the PROSITE database [27], single-amino acid polymorphism from the UniProt variant pages and also information from the GenAge database [41]. The basis for mapping sequence data on structures is an automatic sequence alignment between UniProt and PDB sequences. The PDB sequence is taken directly from the coordinates and not from the SEQRES record.

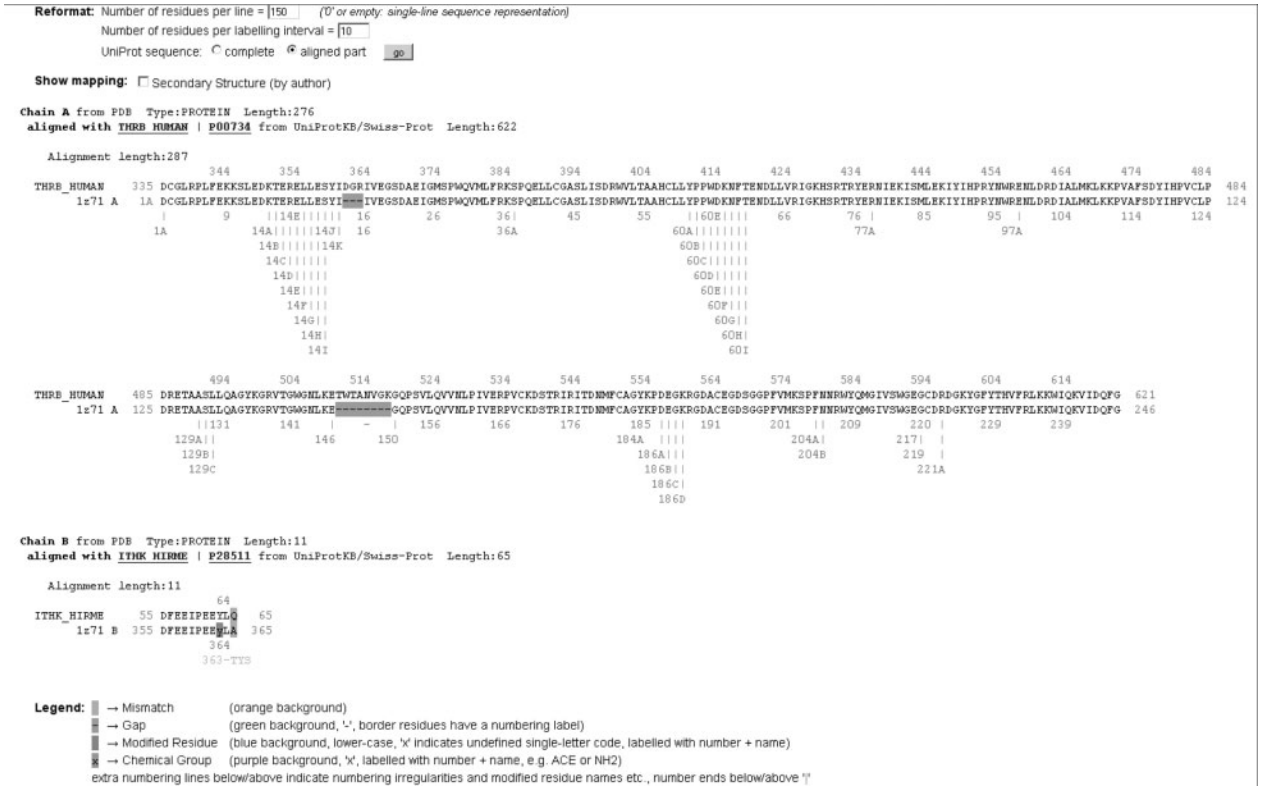

Figure 11: JenaLib UniProt/PDB alignment view for a thrombin/hirudin IIIB complex (1z7I).

The alignment is displayed in an alignment viewer. It highlights mismatches, gaps, modified residues and numbering irregularities in PDB files. The JenaLib alignment can be directly compared to the data shown on the MSD sequence page (Figure 11).

The complete information for a particular structure is shown on one page. There is, however, an expand/collapse mechanism that initially hides the content of most sections (Figure 12) but also allows quick access to the complete information without any additional time-consuming server contacts and tedious clicking through many pages.

Visualization was and is one of the strengths of the JenaLib database. Therefore, it offers a large number of manually generated images. Some of them have been used in newspapers, exhibitions, books and as journal cover images. For example, the journal RNA uses for all issues since 1993 JenaLib molecule representations. The database offers also for all structures automatically generated mono and stereo MolScript images [34] in PDF format. A Virtual Reality Modeling Language viewer can be used for modifying the default orientation. Interactive visualization is possible via RasMol [24] and Chime as well as with the Java-based WebMol [23] and Astex Viewers (www.astex-therapeutics.com/AstexViewer/index.php).

There is also a unique JenaLib viewer based on the platform-independent open-source viewer Jmol (www.jmol.org). The JenaLib Jmol viewer offers a great deal of selection and rendering options by a simple point-and-click mechanism. In addition, there is also a command-line interface to provide access to the Jmol scripting language for advanced users (Figure 13). The viewer gets data from the underlying database mentioned earlier that includes a great deal of information beyond the PDB data. This makes the JenaLib Jmol viewer a unique visualization tool with many options not available in other 3D structure resources. For example, there are a number of standard views that highlight hetero components and sites, PROSITE motifs, SAPs (single amino acid polymorphisms) and the SCOP or CATH domain structure. A unique JenaLib feature is also the rendering of both the CATH and SCOP domain structure within one view. This domain structure can be shown for both asymmetric and biological units.

The standard views can be modified and combined by either a basic or an advanced interface.

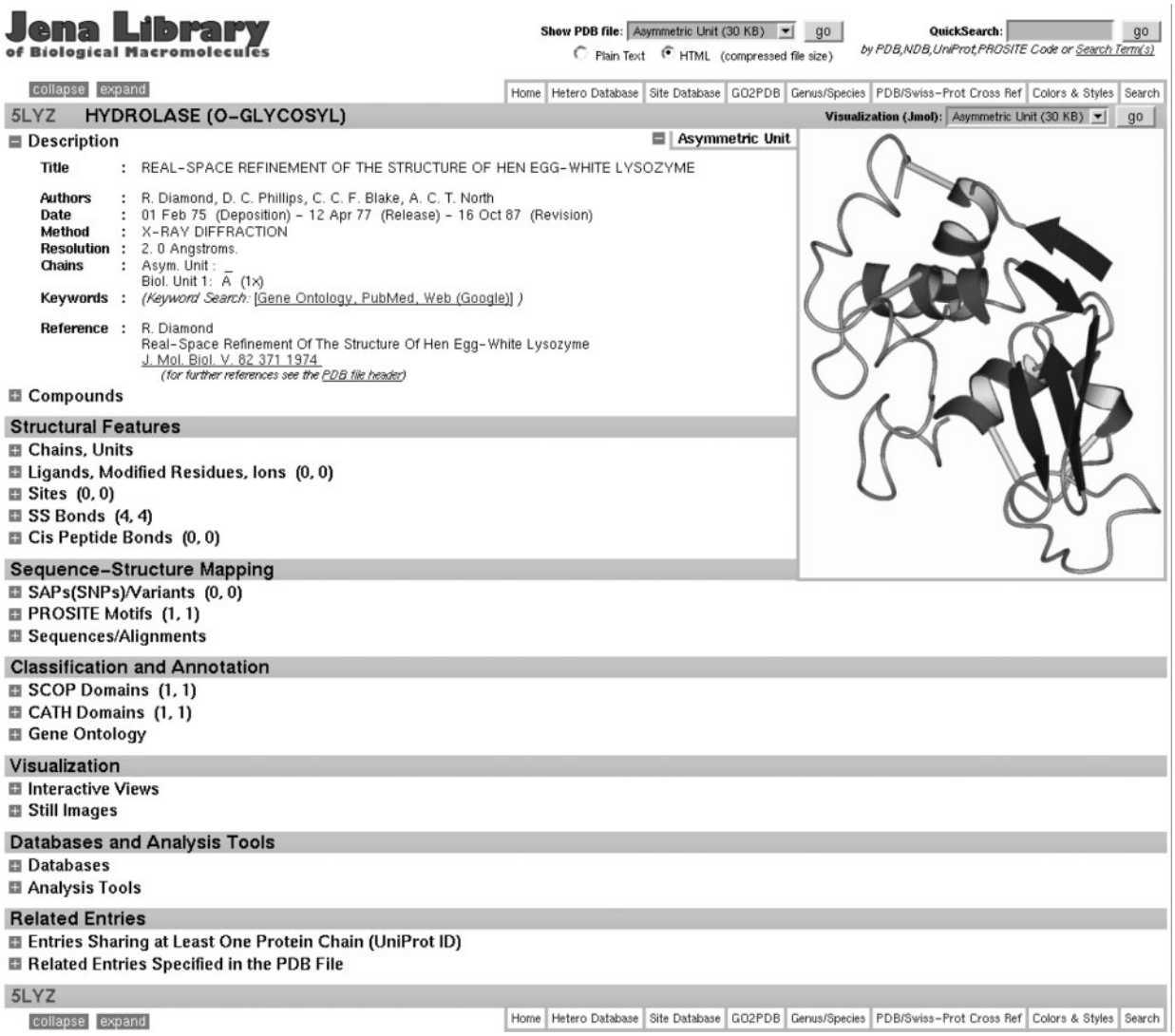

Figure 12: JenaLib atlas page for lysozyme (5 lyz).

In the basic interface, any change in the selection of a pull down menu automatically triggers an action (one-step mechanism). View and selection are coupled in the structure-specific controls. On the other hand, in the advanced interface any change in the selection of a pull down menu only sets the target for the corresponding control buttons (two-step mechanism). View and selection are set independently in the structure-specific controls. There is also an option for exploring individual NMR structure models. The information on biological units is taken from the PDB and not from the PQS server [30].

A further interesting JenaLib option is a content-highlighting system of PDB files which makes the navigation through these files much easier.

The JenaLib offers a number of pre-computed entry lists but also includes an option for the generation of fully customizable lists (Figure 14). Entry sets can be selected according to database (PDB and/or NDB), method (X-Ray/neutron/synchrotron, NMR, electron, other experimental and theoretical model), molecule type (protein, DNA, etc.) or other features such as the occurrence of modified residues, ligands/ions, SAPs and PROSITE motifs as well as availability of information from the SCOP, CATH, OMIM [42] or GenAge databases. The user can select the desired output columns (currently from a list of 26 columns) and define the sort order of the output according to information in any of the columns. The output format is either HTML or ASCII

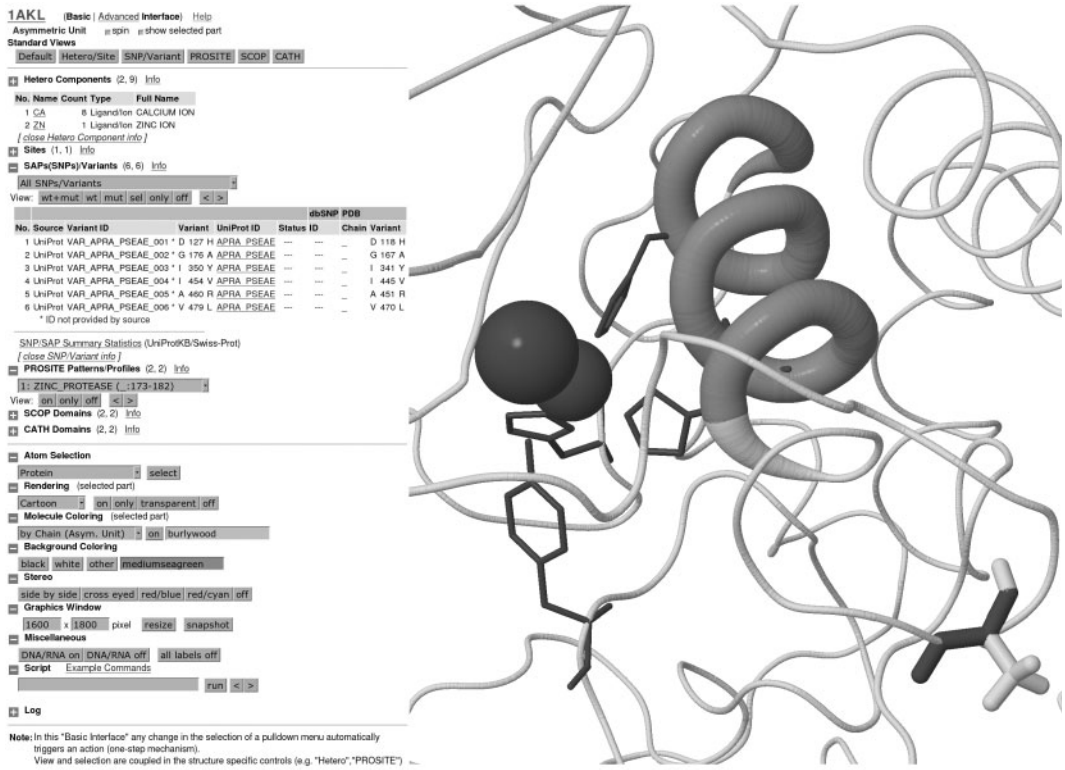

Figure 13: Alkaline protease (1akl) view created with the JenaLib Jmol viewer.Thick ribbon: PROSITE ZINC_PROTEASE motif; large ball: water oxygen atom from the active site (CAT); small ball: Zn; thin sticks: active site amino acids His176, His 180, Tyr216; thick sticks: mutation G167A (wild type ▪ dark, mutant ▪ bright).

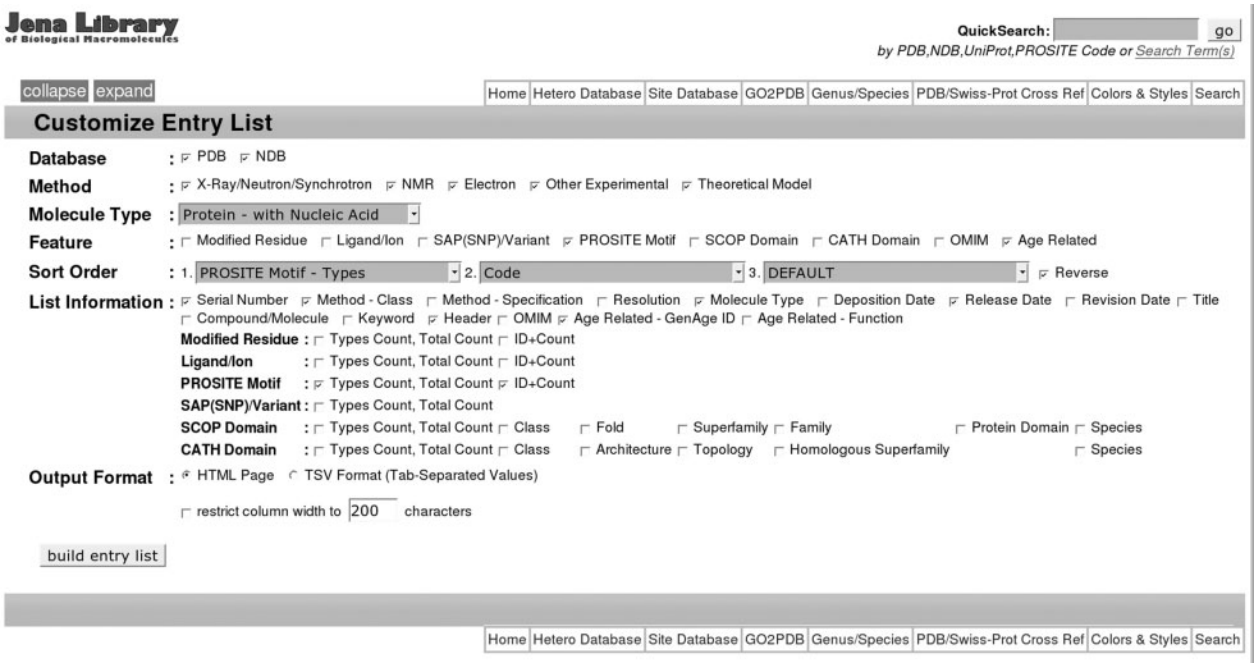

Figure 14: JenaLib entry list customization.

(tab-separated text). Such entry lists can also be generated from QuickSearch results.

QuickSearch is a user-friendly Google-like search option with only one input field. Nevertheless, due to an automatic recognition mechanism, this option allows to search for PDB and NDB IDs, UniProt IDs and accession numbers, PROSITE IDs and accession numbers as well as for other search strings. The search space comprises the database codes mentioned and the HEADER, STRUCTURE, TITLE, KEYWD, EXPDATA, HETNAME and HET, JRNL, COMPND and SOURCE records of the PDB file. The search is also performed in all sub-records such as auth, titl, ref, refn for JRNL or organism_scientific, organism_common, cellular_ location, expression_system, cell_line, tissue for the SOURCE record.

In the hit list, the occurrence of the search terms is indicated with different colours for different terms. This makes the search results really transparent. There is also an advanced search option allowing searches in specific database sections. This option needs to be further improved, however, in order to enable complex queries.

The JenaLib atlas pages contain a large number of linked cross references to a total of 50 external databases and analysis tools with information on a particular structure and using the PDB/NDB ID, UniProt ID/accession number, Enzyme number, OMIM disease ID or GenAge ID for linking. The linked analysis tools include, for example, protein disorder predictors and the PDB cartoon tool.

COMPARING THE RESOURCES

It is difficult for a potential user to identify all the either overlapping or unique features of the 3D structure databases. We have, therefore, compiled Table 1 with compact information on features available in 3D structure databases. An additional Table 2 offers information on a subset of these data with an emphasis on uniqueness.

In addition, a few remarks are to be added. Basically, all resources should have the very same information from the PDB files. This is, however, not completely true because there are currently three different PDB formats, the original ASCII file, XML and mmCIF files. Unfortunately, the data content of these files are not identical. One difference refers to the chain identifiers. In the original PDB files occasionally chain identifiers are not indicated, see for example the lysozyme structure with the PDB ID 5lyz. On the other hand, in the XML and mmCIF all structures have chain identifiers. In the case mentioned this is A. So, the chain information displayed on the atlas page depends on the file type used. In the lysozyme case chain identifier A is given by MSD and OCA, whereas for all other resources no chain identifier is indicated.

Fortunately, there is an announcement of switching to remediated PDB data after July 2007 and by this attempt hopefully the differences currently observed will vanish.

Chain identifiers may also change when passing from asymmetric to biological units. The structure databases use information on biological units from two sources, either directly from the RCSB PDB or from the PQS server. It is important to realize that these sources may provide different results. For example, in the lysozyme case the predicted biological unit is dimeric according to the PQS server, but monomeric according to the PDB.

Search results also depend critically on search space and procedure adopted by the different databases. Extreme differences can be seen, for example, if one tries to identify the number of 'to be published' entries. A search performed on 20 July returned the following results: PDBj Keyword: 7848 entries, RCSB PDB: 8151 entries, MSDlite Text Search: 673 entries, JenaLib QuickSearch: 9877 entries.

Because of these potential differences it is always a good idea to try to get a specific information from different databases. Despite these problems the overwhelming majority of 3D structural data is identical for the resources described.

CONCLUSIONS

Information on 3D structures of biological macro-molecules is available from the primary resource, the RCSB PDB and from a number of secondary resources. In recent years, strong efforts have been undertaken by these resources towards extended data integration, better data uniformity and improved analysis tools with an emphasis on user-friendliness. Even though there is a substantial overlap between these databases, all of them have their unique features. Taken together, the 3D structural resources represent a rich and easily accessible information source that can be used both by the structural

biologist and by the bench biologist. Challenges for the future are the development of improved navigation tools through the increasing amount of structural data and data integration ranging from the genomic level over protein sequences and structures and pathways to disease information.

Key Points

- The primary information resource, the RCSB Protein Data Bank as well as secondary resources represent a rich archive of 3D structural information.
- The rapidly increasing information on 3D structures of biological macro-molecules is not yet fully utilized in functional genomics and proteomics.
- Most of the user interfaces of 3D structure database have now been designed in a really user-friendly manner and can easily be used by the bench biologist.
- This should lead to a stronger impact of 3D structural information of proteins and nucleic acids in the field of functional genomics and proteomics.

Acknowledgements

The article was funded by Nationales Genomforschungsnetz NGFN-2: Grant 01 GR 0457 to J.S. The authors are grateful to Kristina Mehliss who has helped with thumbnail image generation, to Friedrich Haubensak for maintaining the web server and to Stefan Westermeier for working on the QuickSearch option during an internship.

References

1. Bernstein FC, Koetzle TF, Williams GJ, *et al*. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 1977;**112**:535–42.
2. Bourne PE, Westbrook J, Berman HM. The Protein Data Bank and lessons in data management. *Brief Bioinform* 2004;**5**:23–30.
3. Editorial. *Nat Struct Biol* 1997;**4**:329–30.
4. Berman H, Henrick K, Nakamura H, *et al*. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 2007;**35**:D301–3.
5. Doreleijers JF, Mading S, Maziuk D, *et al*. BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *J Biomol NMR* 2003;**26**:139–46.
6. Peitsch MC, Wells TN, Stampf DR, *et al*. The Swiss-3DImage collection and PDB-Browser on the World-Wide Web. *Trends Biochem Sci* 1995;**20**:82–4.
7. Suhnel J. Image library of biological macromolecules. *Comput Appl Biosci* 1996;**12**:227–9.
8. Stampf DR, Felder CE, Sussman JL. PDBBrowse - a graphics interface to the Brookhaven protein data bank. *Nature* 1995;**374**:572–4.
9. Sussman JL, Lin D, Jiang J, *et al*. Protein data bank (PDB): a database of 3D structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr* 1998;**54**: 1078–84.
10. Wu CH, Apweiler R, Bairoch A, *et al*. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* 2006;**34**: D187–91.
11. Laskowski RA, Hutchinson EG, Michie AD, *et al*. PDBsum: a web-based database of summaries and analyses of all PDB structures. *Trends Biochem Sci* 1997;**22**:488–90.
12. Nakamura H, Ito N, Kusunoki M. [Development of PDBj: Advanced database for protein structures]. *Tanpakushitsu Kakusan Koso* 2002;**47**:1097–101.
13. Velankar S, McNeil P, Mittard-Runte V, *et al*. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res* 2005;**33**:D262–5.
14. Berman HM, Westbrook J, Feng Z, *et al*. The Nucleic Acid Database. *Acta Crystallogr D Biol Crystallogr* 2002;**58**: 889–98.
15. Murthy VL, Rose GD. RNABase: an annotated database of RNA structures. *Nucleic Acids Res* 2003;**31**:502–4.
16. Kleywegt GJ, Harris MR, Zou JY, *et al*. The Uppsala Electron-Density Server. *Acta Crystallogr D Biol Crystallogr* 2004;**60**:2240–9.
17. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. *Nature* 1996;**381**:272.
18. Neshich G, Mazoni I, Oliveira SR, *et al*. The Star STING server: a multiplatform environment for protein structure analysis. *Genet Mol Res* 2006;**5**:717–22.
19. Weissig H, Bourne PE. Protein structure resources. *Acta Crystallogr D Biol Crystallogr* 2002;**58**:908–15.
20. Andreeva A, Howorth D, Brenner SE, *et al*. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 2004;**32**:D226–9.
21. Greene LH, Lewis TE, Addou S, *et al*. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res* 2007;**35**:D291–7.
22. Finn RD, Mistry J, Schuster-Bockler B, *et al*. Pfam: clans, web tools and services. *Nucleic Acids Res* 2006;**34**: D247–51.
23. Walther D. WebMol—a Java-based PDB viewer. *Trends Biochem Sci* 1997;**22**:274–5.
24. Sayle RA, Milner-White EJ. RASMOL: biomolecular graphics for all. *Trends Biochem Sci* 1995;**20**:374.
25. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 1997;**18**:2714–23.
26. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004;**32**: D258–61.
27. Hulo N, Bairoch A, Bulliard V, *et al*. The PROSITE database. *Nucleic Acids Res* 2006;**34**:D227–30.
28. Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 2004;**32**: D129–33.
29. Bartlett GJ, Porter CT, Borkakoti N, *et al*. Analysis of catalytic residues in enzyme active sites. *J Mol Biol* 2002;**324**: 105–21.

30. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. *Trends Biochem Sci* 1998;**23**:358–61.
31. Golovin A, Dimitropoulos D, Oldfield T, *et al*. MSDsite: a database search and retrieval system for the analysis and viewing of bound ligands and active sites. *Proteins* 2005; **58**:190–9.
32. Laskowski RA, Chistyakov VV, Thornton JM. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res* 2005;**33**:D266–68.
33. Laskowski RA, MacArthur MW, Moss DS, *et al*. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Cryst* 1993;**26**:283–91.
34. Kraulis PJ. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr* 1991;**24**:946–50.
35. Merritt EA, Bacon DJ. Raster3D: photorealistic molecular graphics. *Methods Enzymol* 1997;**277**:505–24.
36. Luscombe NM, Laskowski RA, Thornton JM. NUCPLOT: a program to generate schematic diagrams of protein-nucleic acid interactions. *Nucleic Acids Res* 1997; **25**:4940–5.
37. Laskowski RA. Enhancing the functional annotation of PDB structures in PDBsum using key figures extracted from the literature. *Bioinformatics* 2007;**23**:1824–7.
38. Wang Y, Addess KJ, Chen J, *et al*. MMDB: annotating protein sequences with Entrez's 3D-structure database. *Nucleic Acids Res* 2007;**35**:D298–300.
39. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. *Curr Opin Struct Biol* 1996;**6**: 377–85.
40. Reichert J, Suhnel J. The IMB Jena Image Library of Biological Macromolecules: 2002 update. *Nucleic Acids Res* 2002;**30**:253–4.
41. de Magalhaes JP, Costa J, Toussaint O. HAGR: the human ageing genomic resources. *Nucleic Acids Res* 2005; **33**:D537–43.
42. McKusick V.A. *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders*. Baltimore, Johns Hopkins University Press, 12th edn, 1998.