

The IMB Jena Image Library of Biological Macromolecules

Jan Reichert, Andreas Jabs, Peter Slickers and Jürgen Sühnel*

Biocomputing, Institut für Molekulare Biotechnologie, Postfach 100813, D-07708 Jena

Received

ABSTRACT

The IMB Jena Image Library of Biological Macromolecules (<http://www.imb-jena.de/IMAGE.html>) is aimed at a better dissemination of information on three-dimensional biopolymer structures with an emphasis on visualization and analysis. It provides access to all structure entries deposited at the Protein Data Bank (PDB) and Nucleic Acid Database (NDB). By combining automatic and manual processing it is possible to keep pace with the rapidly growing number of known biopolymer structures and to provide, for selected entries, information not available from automatic procedures. Each entry page contains basic information on the structure, various visualization and analysis tools as well as links to other databases. The visualization techniques adopted include static mono/stereo raster or vector graphics representations, virtual reality modeling (VRML), RasMol/Chime scripts and Java applets. A helix and bending analysis tool provides consistent information on about 750 DNA and RNA duplex structures. Access to metal-containing PDB entries is possible via the Periodic Table of Elements. Finally, general information on amino acids, *cis*-peptide bonds, structural elements in proteins, base pairs, nucleic acid model conformations, and experimental methods for biopolymer structure determination is provided.

* To whom correspondence should be addressed. Tel: +49 3641 65 6200; Fax: + 49 3641 65 6210; Email: jsuehnel@imb-jena.de

INTRODUCTION

Three-dimensional (3D) structural information on biological macromolecules is an essential requirement for the understanding of biological function and for the variation of this function by rational, evolutionary or combinatorial approaches. The rate at which 3D structures of biopolymers are determined has been dramatically increased within recent years. Currently (October 12, 1999), the Protein Data Bank (1) holds 10858 entries. However, it has been pointed out, that there is only a minor impact of the available 3D biopolymer information on the research communities outside structural biology (2). It was claimed that there may be a rift between the sequence and the structure world due to a clash of scientific cultures and the inherent complexity of structure data. Therefore, tools for a better dissemination of 3D information on biopolymer structures are required.

We have started in 1993 to set up the freely accessible internet based Image Library of Biological Macromolecules with manually generated visual information on selected structure entries (3-5). In 1998 and 1999, the IMB Jena Image Library has been substantially reorganized and extended. Now it provides a visual interface to all currently known biopolymer structures, offers various analysis tools and intends to fulfil both scientific and educational needs.

THE ATLAS OF MACROMOLECULE STRUCTURES

The heart of the data resource is the Atlas of Macromolecule Structures which provides access

to all structure entries deposited at either the Protein Data Bank (PDB) (1) or the Nucleic Acid Database (NDB) (6).

Entry page description

Entry pages are generated automatically for all structures available either from the PDB or from the NDB. In designing these pages it has been our aim to offer as much information as possible in one place. A thumbnail image generated on demand with RasMol (7) shows for proteins a cartoon-like representation. Nucleic acids are displayed as line drawings with a backbone ribbon. Active sites, ligands, modified residues or metal ions and disulfide (SS) bonds are highlighted. In addition to this image, textual information on the occurrence of *cis*-peptide and SS bonds, ligands, modified residues, ions and active site amino acids is given. The reference is linked to the PubMed server. In addition, links to other data resources with information on this entry are provided. Depending on the molecule type they include the following databases and analysis tools: PDB (1), NDB (6), PDBsum/CATH (hierarchical classification of protein domain structures) (8), SCOP (structural classification of proteins) (9), SWISS-PROT (protein sequence database) (10), PRESAGE (structural genomics database) (11), MMDB/Entrez (Entrez 3D structure database) (12), DSSP (dictionary of protein secondary structure) (13), FSSP (fold classification based on structure alignment of proteins) (14), WHAT_CHECK (15), 3DEE (database of protein domain definitions) (16), the ENZYME databank (17), and the Metalloprotein Database at the Scripps Research Institute. The remaining part of the entry pages is devoted to various visualization and analysis tools for the complete structure as well as for ligands, active sites, and *cis*-peptide bonds alone. If

manually generated information on particular entries is available, it is taken into account by the script generating the entry page. The various visualization techniques are described in more detail below. For all protein structures Ramachandran plots can be retrieved from the PDBsum/CATH server and for nucleic acid double helix structures there is a link to special analysis pages on the helix geometry and on bending properties.

Visualization techniques

The visual information provided by the Image Library includes static raster graphics and vector graphics images (mono and stereo), virtual reality modeling (VRML) representations allowing for a limited interactivity, and scripts for visualization within freely available molecular graphics programs. All kinds of visualization rely on the coordinate files, which contain cartesian coordinates of atomic positions and additional information on a structure. From the point of view of information transfer over the web there are currently three different types:

1. Image information is generated either automatically or manually on the database server and is then transferred to the client.
2. Coordinate information is transferred over the web possibly together with additional information for automatic launching of a local program on the client and a special display of the molecular structure.
3. Both the coordinate information and the program code (Java applet) are retrieved from the server.

All three types of information transfer are implemented in the Image Library and each of them has its merits and disadvantages and may be useful for different potential users. According to type 1 the server offers precomputed images and allows for image generation on demand. For the latter technique we use RasMol (7) and MolScript (18). With information from the PDB file a cartoon secondary structure representation of proteins with different colors for helix, loop and sheet elements is generated. Nucleic acids are rendered as backbone and wireframe with different strands shown in different colors. Known active site amino acids in the molecular structure are displayed as sticks, whereas all ligands, including metal ions and coenzymes, are shown as spacefilling models. For structure entries without manually generated images the RasMol representation is used as a thumbnail view on the entry page. MolScript mono and stereo representations are generated as portable document format (PDF) vector graphics files. MolScript is also used for VRML representations. VRML stands for virtual reality modeling language and is essentially a 3D image format supplemented by network functionality (4,19,20). For the visualization of VRML files appropriate viewers are required. They allow the user to zoom, rotate or translate the 3D image objects. As compared to molecular graphics programs they offer a reduced functionality and do not allow to change the display styles, at least in the simple version adopted in the Image Library. The program MolMol is used for semiautomatic image generation (21). Finally, a variety of further molecular graphics programs, like InsightII and WebLab (Molecular Simulations, Inc), Setor (22), Prepi (Imperial Cancer Research Funds), and Midas (23) are applied for the manual generation of images.

Because RasMol is freely available for almost all platforms and also as a browser plugin (Chime, MDL Information Systems, Inc.), it is an appropriate program for information

transfer of type 2 as well. The Image Library provides RasMol scripts which launch RasMol or the corresponding Chime plugin on the client and display the structure in the same manner as described for the thumbnail image. The advantage of this approach is the interactivity offered. The user can select an appropriate point of view as with a VRML viewer, but may, in addition, change rendering modes and color schemes and has a lot of further options for structure analysis.

Examples for information transfer according to type 3 are the links to the WebMol viewer (24) and to a viewer from the Metalloprotein Database hosted at the Scripps Research Institute, which displays metal binding sites. The great advantage of this approach is that all program updates are only necessary on the server and no local installation is required. The disadvantages are concerned with the bandwidth requirements for transferring both data and program code.

Automatic and manual processing

The original idea behind the IMB Jena Image Library was to provide very informative high-quality manually generated images. The advantage of these representations is the high information content. On the other hand, in the light of the rapidly growing number of known structures the fraction of entries with manually generated images as compared to the total number of structures known became smaller and smaller. We have, therefore, adopted an approach which combines automatic and manual processing of entries. The generation of all entry pages and the major part of visualization is done automatically. However, the scripts can take into account manually generated information if for particular entries such information is

available. In this way we keep pace with the fast growth of the number of structures known and, nevertheless, do not neglect the information which cannot be obtained from automatic procedures alone. Finally, there are semi-automatic processing procedures. They reduce the time required for image generation substantially.

Access modes and built-in databases

Access to individual structure entries is possible by search options or via various entry classification schemes. Searching can be done either for the PDB/NDB code or for logical combinations of text strings in both the PDB file headers and the Image Library annotation files. This is probably the main access mode. However, we believe that the access via various structure compilations listed by molecule type and method of structure determination is also useful. The molecule type classification schemes include proteins, nucleic acids, protein-nucleic acid complexes, carbohydrates and RNA structure entries. Further, access is possible via a classification scheme generated by the NDB. Metal-containing entries can be accessed via the Periodic Table of Elements. The IMB Jena Image Library includes a hetero components database. It provides a complete compilation of all hetero components occurring in the PDB entries and compilations of proteins, protein-nucleic acid complexes, nucleic acids and carbohydrates listed by hetero components. In addition, a search option is offered. It allows searches for hetero identifiers and names, chemical elements in the hetero components, and for any character string in the PDB title record. In this way it requires only a few mouse clicks to get a comprehensive information on all protein-nucleic acid complexes obtained

from molecular modeling approaches, for example, or to learn that currently there is only one structure in the PDB which contains the metal ion Ce^{3+} , namely the NMR structure of a calmodulin amino-terminal domain (PDB code: 1ak8) (25).

Analysis tools

Visualization is one way of structure analysis and, therefore, the visualization types described above represent useful analysis tools. Moreover, various structure features, like the occurrence of *cis*-peptide and SS bonds, of ligands and modified residues, and of active sites are displayed on the entry pages in a user-friendly manner. Except for the SS bonds all of them can be visualized separately. Links for the generation of Ramachandran plots are a further analysis option.

The Image Library offers a tool for the analysis of the helix geometry of about 750 nucleic acid double helix structures (free or bound to drugs or proteins) with at least 6 consecutive base pairs. The results are presented on two additional pages. The first page provides quantitative information on the helix geometry comprising tables of the inter-base pair parameters rise, shift, slide, twist, roll, tilt, and of selected backbone torsion angles. In addition, plots of the inter-base pair parameters as well as of the minor and major groove widths can be obtained. Moreover, full outputs of the programs CURVES (26) and FREEHELIX (27) are accessible. Finally, three orthogonal views of the nucleic acid duplex including the helical axis determined with CURVES are shown. The orientation of the duplex structures is not just taken from the PDB or NDB file. Rather, the coordinate axes are aligned

to the principle axes of inertia of the global helix axis. By means of this approach very informative representations are obtained which clearly show the bending features of the helices. Helix bending is a particular important aspect for an understanding of both the sequence-dependent structural variations of free nucleic acids and of nucleic acid-drug and nucleic acid-protein recognition (27). Bending is analyzed in more detail on the second page. Within our analysis approach the helical axis determined with the CURVES algorithm is fitted to the following geometrical models: a straight line, a curved line (arc), a kinked line, and a double kinked line. By means of a goodness-of-fit parameter the most appropriate model can be selected. Using this information and the geometrical parameters of the corresponding models (radius of curvature, kink angles, twist angle) a comprehensive bending classification of nucleic acid duplex structures is possible. Whereas the analysis pages for the nucleic acids are already available, work on the classification is still in progress. A particular strength of this approach is that a consistent analysis procedure can be applied to about 750 nucleic acid structures.

BASIC INFORMATION ON BIOLOGICAL MACROMOLECULES

In addition to the Atlas of Macromolecule Structures, the Image Library contains a division with basic information on the architecture of biopolymers. It is primarily devoted to the non-expert in structural biology and may also be used for educational purposes. By offering both data on the most recent structures solved and on basic principles of the architecture of biological macromolecules it is intended to bridge the gap between information required for educational purposes and for scientific needs. This division includes information on the

following items: experimental methods for biopolymer structure determination, amino acids (The Amino Acid Repository), protein secondary structure elements and on *cis*-peptide bonds (28), base pairs (The Base Pair Directory), nomenclature of nucleic acid structures, and nucleic acid model conformations.

RELATION TO OTHER DATABASES AND ACCESS STATISTICS

The primary resources for 3D biopolymer structure information are the PDB and NDB. Within the last few years they have undertaken substantial efforts to provide structural information in a user-friendly manner. Further databases, which also offer information on 3D structures of biopolymers derived from the structure files maintained at the PDB are PDBsum/CATH (<http://www.biochem.ucl.ac.uk/bsm/pdbsum>), SCOP (<http://scop.mrc-lmb.cam.ac.uk>) and the Molecular Modeling Database (MMDB) (<http://www.ncbi.nlm.nih.gov/Structure/MMDB/>). These databases provide access to all entries of either the PDB or NDB and offer automatically generated images and further information. On the contrary, the SWISS-3DIMAGE collection contains manually generated high-quality images of a few hundred protein structures only, however (29). By combining automatic and manual processing procedures the IMB Jena Image Library represents a combination of both approaches. In addition to the automatically generated image sets for the currently 11000 PDB entries and 800 NDB entries, annotated manually generated image information is available. Currently, the Image Library includes more than 5700 mono and stereo images of this type for 569 structures and slightly more than 4400 images for

approximately 2000 structures obtained in a semiautomatic manner from MolMol. In addition, for about 400 structures distance plots visualizing the distance between representative atoms of all nucleotides and/or amino acids in a structure are available. The annotations highlight particular structural features like ligands, active sites, interaction interfaces and so on.

Users are encouraged to contribute their own images to the Library. One example are images kindly provided by the Prolysis web server which contains information on proteases and protease inhibitors (<http://delphi.phys.univ-tours.fr/Prolysis/>). Moreover, we are willing to provide images on user request. The figures of manually generated images should be compared to the SWISS-3DIMAGE collection (<http://www.expasy.ch/sw3d/>) which offers for 380 structures 640 high-quality images in different file formats (29). The databases described above have strengths in particular fields and therefore the entry pages of the Image Library include links to most of them. Unique features of the Image Library are the great variety of visualization and analysis techniques offered, the combination of automatically and manually generated information, the accessibility of all currently known biopolymer 3D structures within one database, links to many other data resources with information on a particular structure, the various access modes which are not restricted to searches for codes and finally the combination with basic information on biopolymer structures.

Information on the access statistics between June and September 1999 is given in Table 1. The number of entry pages accessed per month varies between 6200 and 25000. It should be noted that the structure entry access numbers may be biased by robot access. Therefore, an additional access statistics is given where each host is counted only once per day. The data in Table 1 also shows that the manually generated images are widely used. Finally, in Table 2 various features of the IMB Jena Image Library are compiled for which a brief introduction on the website is available.

AVAILABILITY AND TECHNICAL REQUIREMENTS

The IMB Jena Image Library of Biological Macromolecules can be freely accessed on the World Wide Web at <http://www.imb-jena.de/IMAGE.html>. To take full advantage of all visualization tools, in addition to a standard web browser like Netscape Communicator or Microsoft Internet Explorer, the local installation of the following free programs/plugins is required: RasMol (7) or Chime (MDL Information Systems, Inc), Adobe Acrobat Reader, a VRML viewer like CosmoPlayer (Silicon Graphics, Inc.), for example. Information on the web addresses for downloading these programs/plugins can be obtained from our various help pages.

ACKNOWLEDGEMENTS

This work is supported by the German Bundesministerium für Wissenschaft, Forschung und Technologie. We are grateful to F. Haubensak, K. Mehliß and C. Schneider for support.

REFERENCES

1. Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,Jr.,E.E., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535-542.
2. Editorial. (1997) *Nature Struct. Biol.*, **4**, 329-330.

3. Sühnel,J. (1996) *Comput. Appl. Biosci.*, **12**, 227-229.
4. Sühnel,J. (1997) In Hofestädt,R., Lengauer,T., Löffler,M. and Schomburg,D. (eds.), *Bioinformatics. Proceedings of the German Conference on Bioinformatics, GCB '96. Lecture Notes in Computer Science.* Springer-Verlag, Berlin, Vol. 1278, pp. 189-198.
5. Sühnel,J. (1997) *Trends Genet.*, **13**, 206-207.
6. Berman,H.M., Olson,W.K., Beveridge,D.L., Westbrook,J., Gelbin,A., Demeny,T., Hsieh,S.H., Srinivasan,A.R. and Schneider,B. (1992) *Biophys. J.*, **63**, 751-759.
7. Sayle,R.A., Milner-White,E.-J. (1995) *Trends Biochem. Sci.*, **20**, 374.
8. Orengo,C.A., Pearl,F.M.G., Bray,J.E., Todd,A.E., Martin,A.C., LoConte,L. and Thornton,J. (1999) *Nucleic Acids Res.*, **27**, 275-279.
9. Hubbard,T.J.P., Ailey,B., Brenner,S.E., Murzin,A.G. and Chothia,C. (1999) *Nucleic Acids Res.*, **7**, 254-256.
10. Bairoch,A. and Apweiler,R. (1999) *Nucleic Acids Res.*, **27**, 49-54.
11. Brenner,S., Barken,D. and Levitt,M. (1999) *Nucleic Acids Res.*, **27**, 251-253.
12. Marchler-Bauer,A., Address,K.J., Chappey,C., Geer,L., Madej,T., Matsuo,Y., Wang,Y. and Bryant, S.H. *Nucleic Acids Res.*, **27**, 240-243.
13. Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577-2637.
14. Holm,L. and Sander,C. (1996) *Science*, **273**, 595-602.
15. Vriend, G. (1990) *J. Mol. Graphics*, **8**, 52-56.
16. Siddiqui,A.S. and Barton,G.J. (1995) *Protein Science*, **4**, 872-884.
17. Bairoch,A. (1999) *Nucleic Acids Res.*, **27**, 310-311.
18. Kraulis, P.J. (1991) *J. Applied. Cryst.*, **24**, 946-950.
19. Vollhardt,H., Henn,C., Moeckel,G. Teschner,M. and Brickmann,J. (1995) *J. Mol. Graphics*, **13**, 368-372.

20. Brickmann,J. and Vollhardt,H. (1996) *Trends Biotechnol.*, **14**, 167.
21. Koradi,R. Billeter,M. and Wüthrich.K. *J. Mol. Graphics*, **14**, 51-55, 29-32.
22. Evans, S.V. (1993) *J. Mol. Graphics*, **11**, 134-138.
23. Ferrin,T.E., Huang,C.C., Jarvis,L.E. and Langridge,R. (1988) *J. Mol. Graphics*, **6**, 13-27,36.
24. Walther,D. (1997) *Trends Biochem. Sci.*, **22**, 274-275.
25. Bentrop,B., Bertini,I., Cremonini,M.A., Forsen,S., Luchinat,C. and Malmendal,A. (1997) *Biochemistry*, **30**, 11605-11618.
26. Lavery,R. and Sklenar,H. (1988) *J. Biomol. Struct. Dyn.*, **6**, 63-91.
27. Dickerson,R.E. (1998) *Nucleic Acids Res.*, (1999) **26**, 1906-1926.
28. Jabs,A., Weiss,M.S. and Hilgenfeld.R. (1999) *J. Mol. Biol.*, **286**, 291-304.
29. Peitsch,M.C.,Stampf,D.,R.,Well,T.N.C. and Sussman,J.L. (1995) *Trends Biochem. Sci.* **20**, 82-84.

Table 1. Access statistics of the IMB Jena Image Library of Biological Macromolecules for June to September 1999 (outside users only)

| <i>Month</i> | <i>June</i> | <i>July</i> | <i>August</i> | <i>September</i> |
|--|-------------|-------------|---------------|------------------|
| Access to structure entries | 16969 | 10122 | 6284 | 25293 |
| Hosts ^{a)} | 2243 | 1120 | 989 | 1192 |
| Number of retrieved manually generated image files | 16587 | 6695 | 3284 | 4859 |

a) Each host is counted once per day.

Table 2. Compilation of features of the IMB Jena Image Library of Biological Macromolecules for which supplementary material is available at http://www.imb-jena.de/IMAGE_SUPPL.html.

- Entry page with automatically generated information only
- Entry page with additional manually generated information
- Access modes
- Helix and bending analysis of nucleic acid duplex structures
- The Base Pair Directory
- The Amino Acid Repository