Proceedings of the Workshop “Multimedia-Systeme für Wissenschaft und Technik”, Magdeburg, 22. September 1998, H.-J. Appelrath, D. Boles, K. Meyer-Wegener (Hrsg.)

Views of Biological Macromolecules on the World-Wide Web: The IMB Jena Image Library

Jan Reichert, Andreas Jabs, Jürgen Sühnel
Biocomputing, Institut für Molekulare Biotechnologie
Postfach 100813, D-07708 Jena
E-Mail: {jr,ajabs,jsuehnel}@imb-jena.de

Abstract

The IMB Jena Image Library of Biological Macromolecules *(http://www.imb-jena.de/IMAGE.html)* is a freely accessible Internet archive with three-dimensional (3D) structural information on proteins, nucleic acids and carbohydrates and an emphasis on visualization. The visualization techniques used include static mono and stereo representations, virtual reality modeling, the usage of public domain molecular graphics programs, which require a local installation, and Java applets where both the program and the underlying coordinate file are transferred over the web. By combining automatic and manual processing it is possible to keep pace with the rapidly growing number of known biopolymer structures and, nevertheless, to provide, at least for selected entries, information which cannot be obtained from automatic procedures. To complement visual information each structure entry contains also links to other structural, sequence and bibliographic databases. In addition, the Library has a division with information on basic principles of biopolymer architecture. The Image Library is a visual interface to 3D biopolymer structures which intends to fulfill both scientific and educational needs.

1 Introduction

Structural information on biological macromolecules is an essential requirement for our understanding of biological function and for a deliberate variation of this function by rational, evolutionary or combinatorial approaches. Progress in recombinant DNA technology and RNA synthesis, X-ray crystallography and nuclear magnetic resonance (NMR) instrumentation and computer and software technology has led to an increasing rate of accumulation of new structures. The archives for depositing three-dimensional (3D) structural information on biopolymers at atomic resolution are the Protein Data Bank (PDB) and the Nucleic Acid Database (NDB) [BKW77, BOB92]. Currently (August 5, 1998), the Protein Data Bank at the Brookhaven National Laboratory has coordinate entries of 7457 proteins and of 594 nucleic acids. In 1997 1640 new structures have been released which corresponds to a growth rate of more than four new structures per day. Together with data generated by sequencing efforts, especially in the various genome projects, by high-

throughput screening adopting combinatorial and evolutionary approaches and in other fields this has led to a data explosion in current biology. In face of these developments and challenges a new discipline has emerged, bioinformatics, which covers all aspects of acquisition, processing, storage, distribution, analysis, and interpretation of biological data [Ben96, Lya96, Rue96].

Simultaneously, the development of new network information systems has led to a phenomenal growth of the world-wide computer network Internet [SH94]. In fact, without the web it would be absolutely impossible to make accessible to the scientific community the huge amounts of data generated by current biomedical research.

Recently, 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. 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 [Edi97]. Therefore, tools for a better dissemination of 3D information on biopolymer structures are required. It is immediately obvious that visualization plays a central role for such attempts.

Since the early 1980s interactive computer graphics has greatly facilitated and improved the visualization of biopolymer structures. The usual approach is to retrieve the coordinate files from a structure database and then to use one of the molecular graphics software packages. This is the method of choice for an in-depth analysis of biopolymer structures. On the other hand, one would often prefer to have biopolymer images directly available without the need to spend some time for visualization or even without having access to a molecular graphics software. As already mentioned this is especially important for the large and heterogeneous community outside structural biology. Already in 1993 we have, therefore, started to set up an Internet-based Image Library of Biological Macromolecules [Süh96, Süh97a, Süh97b, RJS98]. Meanwhile, the major biopolymer structure resources, the Protein Data Bank and the Nucleic Acid Database, have undertaken substantial efforts to improve the quality and user-friendliness of accessing their databases. Nevertheless, additional information resources can play an important role for a better dissemination of structure information on biopolymers.

2 Database organization

The IMB Jena Image Library of Biological Macromolecules was set up in 1993 as a simple directory tree containing an annotation text file and a variety of image files for each entry and a gopher-based access. With the rapid acceptance of hypertext browsers the hypertext transfer protocol (http) has been supported by the Image Library.

Currently, the Library consists of two major divisions (Fig. 1). The heart of the data resource is the access to the biopolymer structure entries. However, from the very beginning it has been our intention to offer also general information on basic principles of biopolymer architecture. Therefore, the Image Library contains in addition information on amino acids (Amino Acid Repository), nucleotides and base pairs, on secondary structure elements of proteins and on DNA model conformations. Finally, still more general structural biology information is available. It includes currently information on structural biology nobel prizes and on experimental methods for the determination of three-dimensional biopolymer structures.

Due to its evolution history the backbone of the Image Library consists of a directory tree with the individual structure entries, which contain manually generated images and annotation files, in

Image Library
of
Biological Macromolecules

JENA

The IMB Jena Image Library of Biological Macromolecules contains visual and other information on three-dimensional biopolymer structures. It provides access to all structure entries deposited at the Protein Data Bank (PDB) including nucleic acid-containing X-ray entries also available at the Nucleic Acid Database (NDB). In addition, general information on the architecture of biopolymer structures is available. The major part of entries has automatically generated images and interfaces for public domain molecular graphics programs. However, the number of manually generated high-quality and informative mono, stereo and VRML represention is steadily increased and remains one of the peculiarities of the IMB Jena Image Library.

- **General Information on Biological Macromolecules**
  - Experimental determination of biopolymer structures
  - Structural biology Nobel prizes
  - Read the Watson Crick paper: A structure for Deoxyribose Nucleic Acid
  - The Amino Acid Repository
  - Structure elements of proteins: helix, beta strand, extended conformation
  - Cis-peptide bonds in proteins
  - DNA model conformations
  - Deoxy-ribonucleotides and Watson-Crick base pairs
  - Definition of strand direction and torsional angles in nucleic acids
  - Standard and modified ribonucleotides from transfer RNA
- **Macromolecule Structures**

  Access to the IMB Jena Image Library via
  - PDB code
  - Search of the PDB file header and of the Image Library annotations
  - Index of all PDB entries (more than 1 Mbyte !!)
  - Index of all PDB **protein entries** sorted according to the method of structure determination
  - Index of all PDB **nucleic acid entries** (no protein-nucleic acid complexes) sorted according to the method of structure determination
  - Index of all PDB **protein-nucleic acid entries** sorted according to the method of structure determination
  - Index of all PDB **carbohydrate entries** sorted according to the method of structure determination
  - Index of all PDB **diffraction entries** sorted according to the molecule type
  - Index of all PDB **NMR entries** sorted according to the molecule type
  - Index of all PDB **model entries** sorted according to the molecule type
  - Index of all PDB **RNA entries**
  - Index of **nucleic acid-containing entries** available in both the NDB and PDB via the NDB Archives classification (X-ray structures only)

Since 1998 the Image Library is supported by the bmb+f Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie

Figure. 1. Homepage of the IMB Jena Image Library of Biological Macromolecules

its leaves. Several indices are built to simplify automated querying and entry generation. They include a list for the content of the whole PDB, a list of all manually generated entries of the IMB Jena Image Library of Biological Macromolecules and an index of the HTML annotation files associated with the images. These indexes are updated in daily or weekly intervals.

Access to individual structure entries is either possible by search options or via various entry classification schemes. Searching can be done either for the PDB entry code or for logical combinations of text strings in both the PDB file headers and the Image Library annotation files. The search in the PDB file headers is performed on one of the PDB mirror sites with the PDB 3DB Browser™ (Jaime Prilusky, 1996-1998). Simultaneously, the Image Library HTML annotation files are processed locally using SWISH (Simple Web Indexing System for Humans, Copyright © 1994, 1995, Enterprise Integration Technologies). The classification schemes include indices of protein, nucleic acid, protein-nucleic acid and carbohydrate entries sorted according to the method of structure determination. Furthermore, indices of diffraction, NMR, and model entries sorted according to the molecule type are available. All these indices are generated automatically from files provided by the PDB or NDB.

3 Automatic versus manual processing

The original idea behind the IMB Jena Image Library was to provide very informative high-quality manually generated images. A similar approach has been chosen simultaneously by the Swiss-3D-Image database [PSW95]. The advantage of these representations is the high information content. The viewpoint can be carefully selected. Various rendering techniques can be combined to provide an optimal impression of the molecular structure. Manually added annotations can assist the user in understanding the structural information and it is simply possible to generate detailed views of especially interesting parts of the structure, for example, of the nucleic acid-protein interaction region in DNA-protein complexes.

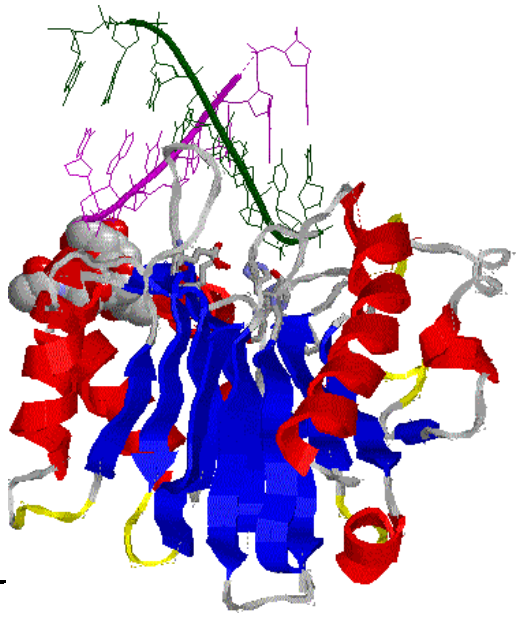
On the other hand, in the light of the rapidly growing number of known structures the fraction of entries with manually generated images as compared to the total number of structures known becomes smaller and smaller. We have, therefore, completely reorganized and extended the database. An approach has been adopted which combines automatic and manual processing of entries. In this way we keep pace with the fast growth of number of structures known and, nevertheless, do not neglect, at least for selected entries, the information which cannot be obtained from automatic procedures alone.

Information on biopolymer structures provided by the IMB Jena Image Library is derived from structure files of the Protein Data Bank which includes protein, nucleic acid and carbohydrate structures determined with different experimental techniques, like diffraction methods and NMR spectroscopy, or by modeling procedures. On the contrary, the Nucleic Acid Database holds so far only nucleic acid containing structures determined by X-ray crystallography. We are maintaining a local version of the PDB which is updated daily. An automatic script generates HTML pages for each structure entry on request (Fig. 2). This page contains basic structural and bibliographic information (type of structure determination, release date, ligands, active site, ...), automatically generated static mono and stereo images and a VRML representation, and links to other databases with information on this entry. Finally, generally available visualization programs or browser plugins can be launched. In this way the IMB Image Library offers an up-to-date access to all currently known biopolymer structures. In addition, for a variety of structures manually generated visual information is provided. To reduce the time required for manual image generation we are attempting to write scripts for a

Image Library
of
Biological Macromolecules

ENDONUCLEASE

DEOXYRIBONUCLEASE I (DNASE I) (E.C.3.1.21.1) COMPLEXED WITH DNA (5'-D(*GP*GP*TP*AP*TP*AP*CP*CP)-3')



PDB code: 1DNK
Authors: S. A. Weston, A. Lahm, D. Suck
Date: 10 Aug 92
Resolution: 2. 3 Angstroms.
Site: ACT: *Glu 39 , Gly 78 , His 134 , Asp 212 , His 252*
Ligands (including metals): N-Acetyl-D-Glucosamine , N-Acetyl-D-Glucosamine
Reference: S. A. Weston, A. Lahm, D. Suck
The X-Ray Structure Of The Dnase I-D(Ggtatacc)2 Complex At 2. 3 Angstroms Resolution
J. Mol. Biol. V. 226 1237 1992

PDB header

Other databases with information on this entry:

NDB | PDB | PDBsum/CATH | SCOP | Entrez/MMDB | DSSP | 3Dee | GRASS | STING | ENZYME

Visualization

Molecular Graphics Programs (visualization and analysis)

- RasMol/Chime (requires local installation; see help)
- WebMol (requires no local installation; loading may be slow)

Structure Images (Gif, PDF, VRML format)

The images are generated automatically by RASMOL (thumbnail), MOLSCRIPT or in a semiautomatic manner by MOLMOL. For particular entries manually generated images from INSIGHTII, MIDAS, SETOR, SYBYL, WEBLAB are available. In almost all cases this standard coloring scheme is used. If you want to retrieve VRML representations read this.

MolScript

- **protein, nucleic acid: cartoon; ligands: spacefill; active site: sticks (PDF: mono, stereo) (VRML2)** (automatic)

MolMol

- **protein, nucleic acid: cartoon; ligands: spacefill; active site: sticks (mono, stereo)**

- Would you like to have an image generated? Let us know.
- Would you like to contribute your own images to the IMB Jena Image Library ? Please mail us.

Figure 2. Entry page for a DNA−protein complex (DNaseI complexed with DNA; PDB code: 1dnk; [WLS92]). This is an entry for which only automatically generated images are available.

semi-automatic processing. The total number of manually generated image files is already larger than 4000. Users are encouraged to contribute their own images to the Library. Moreover, we are willing to provide manually generated images on user request.

4 Visualization techniques

All types of visualization rely on the basic coordinate files, which contain cartesian coordinates of atomic positions. From the point of view of information transfer over the web there are currently three different approaches:

1. Image information can be generated either automatically or manually on the database server and is then transferred to the user.
2. Coordinate information is transferred over the net possibly together with additional information for automatic launching of a local program on the user computer and a special display of the molecular structure.
3. Both the coordinate information and the program are transferred between database server and user.

All three types of information transfer are implemented in the Image Library and each of them has its merits and disadvantages. Variant 1 enables one to provide high-quality rendered images which require software not generally available. For manual high-quality structure rendering the following programs were used: InsightII, WebLab (Molecular Simulations, Inc.), MIDAS [FHJ88], SETOR [Eva93], SYBYL (Tripos Associates, Inc.). For the program MolMol [KBW96] a script for semi-automatic processing was written. It reduces the time required for image generation substantially but nevertheless allows for a restricted interactivity. For automated image generation on the fly a program is required for which the launching, image generation and storing processes are fast. We found RasMol [SM95] to be an ideal rendering program for this purpose, although it does not generate high-quality rendered images. With information from the PDB file RasMol generates a cartoon secondary structure representation of proteins with different colors for helix, loop and sheet. 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 shown as sticks and all ligands, including metal ions and coenzymes, for example, are shown as spacefilling models (Fig. 2). For each structure entry for which no manually generated images are available this representation is used as a thumbnail view on the entry page. Very recently a new version of the program MolScript has become available [Kra91]. It also fulfills the needs required for automatic image generation. Therefore, it is used, in addition to RasMol, for image generation on demand. In this case mono and stereo static PDF (Portable Document Format) images and a VRML representation are provided.

VRML stands for virtual reality modeling language. It is essentially a 3D image format supplemented by network functionality [VHM95, BV96, RJS98]. For the visualization of VRML files corresponding viewers are required. They allow the user to interact with the 3D image objects to a certain extent (zoom, rotation, translation). The first of these viewers have become available in 1995. We have already in 1995 included VRML representations in the Image Library. This was one of the first VRML applications in biology. As compared to molecular graphics programs they offer a reduced interactivity. Nevertheless, there is an advantage over static images. Recently, a second generation VRML format VRML-2.0 has been defined. It offers extended possibilities for interactivity. What is especially important for high-quality representations, however, concerns the reduced filesize. Within VRML-1.0 high-quality representations of large structures may lead to very large files [RJS98]. The filesize can be substantially reduced in VRML-2.0. Experiments with the new

version of the program MolScript [Kra91], which generates VRML-2.0 representations, show very encouraging results.

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 variant 2 as well. The Image Library provides RasMol scripts which activate RasMol or the corresponding Chime plugin on the user computer and display the structure. The advantage of this approach is the interactivity offered. The user can select an appropriate point of view, change rendering modes and color schemes and has a lot of further options.

Finally, the Image Library offers the possibility to invoke the WebMol viewer [Wal97]. This is a Java-based applet where both the coordinate files and the program are transferred over the web. The great advantage of this approach is that all program updates are only necessary on the server. Moreover, it requires no local installation. This, was, for example, greatly appreciated by one of the authors during a recent conference where the local library provided computer access for conference participants but had no RasMol installation. The disadvantages are concerned with the bandwidth requirements for transferring both data and program.

5 Integrating other databases

The value of a database can be greatly increased if it includes links to other databases with related information. For each entry, we, therefore, offer links to the following information resources: PDB [BKW77], NDB [BOB92], PDBsum/CATH [OMJ97], SCOP [HMB97], SWISS-PROT [BA98], enzyme databank [BA96], ENTREZ [Mce98], HSSP [DSS98] FSSP [HS98]. In addition, the PubMed bibliographic database is linked in.

6 Response

The access statistics provides information on the usage of the Image Library. Between October 1997 and February 1998 there were on the average 3000 users accessing the Image Library start page per month. The average number of files retrieved (html, gif, wrl, pdf) was about 45000 per month. The IMB Jena Image Library has been featured as webpick of the day on Februray 20, 1997 by the ‘webzine’ HMS Beagle and has been selected for the Scout report on September 5, 1997. In November 1997 it was described in the Webwatch section of Science (Science 1997, 278, 1089) and it has been chosen as a three star "indispensible" site by the BioMedLink database in July, 1998. Finally, images and animations from the IMB Jena Image Library have also been used in the exhibits 'Gene Worlds: Workshop Man?' of the Deutsches Hygiene-Museum Dresden (March 27, 1998 - January 10, 1999) and ‘ZoomIN/ZoomOut’ organized by the Laboratorio dell’Immaginario Scientifico, Trieste (March 23-29, 1998).

7 Conclusions and outlook

The IMB Jena Image Library offers a ‘visual interface’ to 3D structure information on biopolymers. It hopefully contributes to a better dissemination of information on biopolymer structures. By offering both data on the most recent structures solved and on basic principles of the architecture of biological macromolecules it is intended to lower the barrier between information required for educational

purposes and for scientific needs. A major challenge for future work is the improvement of the information content by automatic or semi-automatic image generation. Our aim is further to extend the part on principles of macromolecule structures, which is very often requested from the Image Library. Moreover, it is intended to include animations of dynamic processes. Finally, we want to improve the scientific value of the Library by including results of a more sophisticated analysis of PDB coordinate files. Work already in progress is related to the occurrence of cis-peptide bonds in proteins and an oligonucleotide bending database.

References

[Bai96] A. Bairoch. The ENZYME databank in 1995. Nucl. Acids. Res. 24:221-222, 1996.

[BA98] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucl. Acids. Res. 26:38-42, 1998.

[Ben96] D. Benton, D. Bioinformatics - principles and potential of a new multidisciplinary tool. Trends Biotechnol. 14:261-273, 1996.

[BKW77] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi. The Protein Data Bank: a computer-based archival file for macromolecule structures. J. Mol. Biol. 112:535-542, 1977.

[BOB92] H. M. Berman, W. K. Olson, D. L. Beveridge, J. Westbrock, A. Gelbin, T. Demeny, S. H. Hsieh, A. R. Srinivasan, B. Schneider. The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys. J. 63:751-759, 1992.

[BV96] J. Brickmann, H. Vollhardt. Virtual reality modeling on the World-Wide Web: a paradigm shift in molecular modeling ? Trends Biotechnol. 14:167, 1996.

[DSS98] C. Dodge, R. Schneider, C. Sander. Nucleic Acids Res. 26:313-315, 1998.

[Edi97] Editorial. Structure and the genome. Nature Struct. Biol. 4:329-330, 1997.

[Eva93] S. V. Evans. (1993) SETOR: a hardware-lighted three-dimensional solid model representation of macromolecules. J. Mol. Graph. 11:134-138, 1993.

[FHJ88] T. E. Ferrin, C. C. Huang, L. E. Jarvis, R. Langridge. The MIDAS display system. J. Mol. Graphics 6:13-27,36, 1988.

[Hal95] S. Hall. Protein images update natural history. Science 267:620-624, 1995.

[HS98] L. Holm, C. Sander. Touring protein fold space with Dali/FSSP. Nucleic Acids Res. 26:316-319, 1998.

[KBW96] R. Koradi, M. Billeter, K. Wüthrich. MOLMOL: a program for display and analysis of macromolecular structures. J. Mol. Graph. 14:51-55, 1996.

[HMB97] T. J. P. Hubbard, A. G. Murzin, S. E. Brenner, C. Chothia. SCOP: a structural classification of proteins database. Nucl. Acids Res. 25:236-239, 1997

[Kra91] P. J. Kraulis. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. J. Applied. Cryst. 24:946-950, 1991.

[Lya96] A. Lyall. Bioinformatics in the pharmaceutical industry. Trends Biotechnol. 14:308-312, 1996.

[Mce98] J. McEntyre. Linking up with ENTREZ. Trends. Genet. 14:39-40, 1998.

[OMJ97] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, J. M. Thornton. CATH - a hierarchical classificationof protein domain structures. Structure 5:1093-1108, 1997.

[PSW95] M. C. Peitsch, D. R. Stampf, T. N. C. Wells, J. Sussman. The Swiss-3DImage collection and PDB-Browser on the World-Wide Web. Trends Biochem. Sci. 20:82-84, 1995.

[RJS98] J. Reichert, A. Jabs, J. Sühnel. Virtual reality modeling. Nachr. Chem. Tech. Lab. 46:A64-A68, 1998.

[Rue96] N. Ruediger. Bioinformatics: New frontier call young scientists. Science 273:265, 1996.

[SH94] B. R. Schatz, J. B. Hardin. NCSA Mosaic and the World Wide Web: Global hypermedia protocols for the Internet. Science 265:895-901, 1994.

[SM95] R. Sayle, E.-J. Milner-White. RASMOL: biomolecular graphics for all. Trends Biochem. Sci. 20:374, 1995.

[Süh96] J. Sühnel. Image Library of Biological Macromolecules. Comput. Appl. Biosci. 12:227-229, 1996.

[Süh97a] J. Sühnel. (1997) Virtual Reality Modeling for Structural Biology. In: Bioinformatics. Proceedings of the German Conference on Bioinformatics, GCB '96. Lecture notes in computer science Vol. 1278, (R. Hofestädt, T. Lengauer, M. Löffler, D. Schomburg, eds.) Springer-Verlag Berlin, 189-198, 1997.

[Süh97b] J. Sühnel. Views of RNA on the World-Wide Web. Trends Genet. 13:206-207, 1997.

[VHM95] H. Vollhardt, C. Henn, G. Moeckel, M. Teschner, J. Brickmann. Virtual reality modeling language in chemistry. J. Mol. Graph. 13:368-372, 1995.

[Wal97] D. Walther. WebMol - a Java-based PDB browser. Trends Biochem. Sci. 22:274-275, 1997.

[WLS92] S. A. Weston, A. Lahm, D. Suck. (1992) The X-ray structure of DNase I-d(GGTATACC)2 at 2.3 Å resolution. J. Mol. Biol. 226:1237-1256, 1992.