A dynamic look at structures: WWW-Entrez and the Molecular Modeling Database

Christopher W.V. Hogue, Hitomi Ohkawa, and Stephen H. Bryant

Trends in Biochemical Sciences 21: 226-229

(reproduced with permission)

Only a few short years ago, accessing biomolecular structures from the Brookhaven Protein Databank (PDB) database involved purchasing a set of magnetic tapes designed for a mainframe computer. Stuctures, and the critical information they contain, were out of reach to anyone with a personal computer. By how quickly things change. Nowadays, the World Wide Web (WWW) has brought forth several excellent services for freely obtaining any file from the PDB structure database (Box 1). Also, there are two excellent viewing programs, MAGE [1, 2], and RASMOL [3], both of which have been made freeely available by their authors for use with IBM, Macintosh and Unix systems. Kinemage has a special niche as it allows extensive annotation and description to be carried along with the structure data, making it a useful resource for teaching [4] . (The journal *Protein Science* has a WWW site containing a remarkable collection of Kinemage format structure files, each created individually by experts and all freely available; see Box 1 for URL.)

Box 1. World Wide Web molecular structure resources

Databases and information

- World Wide Web (WWW)-Entrez and Molecular Modeling Database (MMDB) access.
  http://www.ncbi.nlm.nih.gov
- MMDB Information
  http://www.ncbi.nlm.nih.gov/Structure
- Protein Science Kinemages
  http://www.prosci.uci.edu/kinemages/KinemageIndex.html
- RASMOL home page
  http://www.umass.edu/microbio/rasmol
- Brookhaven Protein Databank (PDB)
  http://www.pdb.bnl.gov
- Other three-dimensional services
  http://cmm.info.nih.gov/modeling/net_services.html
- Chemical Multipurpose Internet Mail Extension (MIME-types)
  http://www.ch.ic.ac.uk/

Software

- RASMOL ftp://www.dcs.ed.ac.uk/generated/package-links/rasmol
- Mage ftp://suna.biochem.duke.edu/pub
- Cn3D http://www.ncbi.nlm.nih.gov/Structure/cn3d.html

Molecular Modeling Database

The Molecular Modeling Database (MMDB) is the new structural division of Entrez [5], provided by the National Center for Biotechnology Information (NCBI). It holds all the structures in the PDB database, but in a different file format [specified in the Abstract Syntax Notation 1 (ASN.1) data description language [6]]. This format allows files of structural data to be readily compressed and exchanged between modern computers. Our hope is that with this new transformed data, structural scientists can start to design and create sofware tools that allow all of us to see different kinds of data, such as structural superpositions and non-atomic three-dimensional models from electron microscopy [7] in a single viewing environment.

The translation of PDB data files into ASN.1 has involved the use of a sophistocated PDB file parser*, which was originally developed as part of an ongoing research project into predicting protein structure from sequence data [8,9]. This parser can 'sort' specific pieces of data from a multitude of information, and furthermore can detect and correct a variety of ambiguities that arise in the PDB file format.

* A parser is a program that extracts the required data items from a text file, from among hundreds of different data locations, containing anything from three-dimensional coordinate values, molecular sequences, to the names of the authors in journal citations.

One common example of such an ambiguity can be seen by viewing almost any structure data file containing the prosthetic groups NAD or FAD (e.g. 1LVL or 1TDE) with RASMOL. These prosthetic groups are colored pink for no apparent reason under the CPK coloring scheme, which colors atoms by their element type.

However the reason for the miscoloration is a conflict between the names given to certain atoms in the PDB file, and the International Union of Pure and Applied Chemistry (IUPAC) symbols for elements; these reside in the same column in the data file. For example, the NAD or FAD prosthetic groups often use names like 'AC1' to abbreviate 'adenine carbon' which is in conflict with the symbol for the element actinium (Ac). The PDB-file format does not allow lowercase letters and so cannot distinguish between the two possibilities. RASMOL is therefore left with no choice but to color these atoms all pink, as it would for other heavy atoms. For RASMOL, the results are a curiosity, but for software that examines mutants of the docking sites near these FAD or NAD groups, this kind of ambiguity could easily cause a serious error.

The MMDB parser also creates explicit bond information for each molecule, called a 'chemical graph', which is recorded in the ASN.1 file. The bond information is obtained, in part, by comparing the parsed data against a dictionary of standard amino acid and nucleic acid residue bonding, naming and connectivity. The inclusion of the chemical graph data can allow a more easily interpreted drawing to be made of some structures (see Fig. 1). This behavior is again a curiosity in graphics programs, but the failure to identify residue types, element types or bonds using modeling programs spells disaster.

Figure 1a. The nuclear magnetic resonance (NMR) average of parvalbumin with bound Ca(2+) (pink spheres) from the structure file 3PAT [10] and displayed using the RASMOL command line. The backbone is drawn using an alpha-carbon trace and the alpha-helices are highlighted in red. All the phenylalanine sidechains are colored blue. Six out of nine of the phenylalanine residues have extra bonds drawn across the aromatic rings. This is owing to the sidechain rotation that appears in the models, which were averaged around the plane of the ring (the beta-delta bond rotation) and ultimately averages into a 'thin' ring.

Figure 1b. The same structure shown from MAGE obtained from an Entrez/Molecular Modeling Database (MMDB) genterated Kinemage using the 'color by Structure' style. The backbone is an alpha-carbon trace or 'virtual' backbone, and its helices are highlighted in blue. The Ca(2+) ions are included and colored automatically in the default Kinemage. Note that the phenylalanine side chains are colored purple and have no extraneous bonds.

Database interconnection

Perhaps more important than these subtle changes, the creation of the MMDB database has allowed the PDB structure data to be connected with both the GenBank and MEDLINE databases. The Entrez sequence databases now contain all the sequences from the PDB structure database, thanks to the MMDB parser, and the bibliographic information from the PDB is parsed and matched to corresponding MEDLINE entries.

Searching a structure database using the Internet is nothing new, but from within WWW-Entrez, one can directly ask the question: are there any structures related to this sequence? To do this, the user starts by retrieving a sequence of interest from the Entrez nucleotide or protein sequnce databases. A list of sequence neighbors [5], which have already been precomputed at the NCBI using the sequence search tool BLAST [11], is asked for next. Finally, links into the structure database will produce a list of three-dimensional structure 'hits', if any exist. A structural biologist may take a different approach, for example, starting with a sequence

and linking back to find related sequences.

At any point, whether searching structures, sequences or their neighbors, the corresponding MEDLINE abstracts can be examined. Many related abstracts can often be found in Entrez, through the MEDLINE text neighboring scheme, some of which can be more recent than the structure file itself. This allows the reader to form a rapid appreciation for the biology related to a structure.

Structure services from WWW-Entrez

For now, WWW users can set up their browser software to automatically start RASMOL or Kinemage from any of the atomic-resolution structure databases using Multipurpose Internet Mail Extensions (MIME-types: Box 1.). Once this facility is set up, any of the structure database services mentioned here can be used as sources for browsing atomic-resolution three-dimensional structures.

Small bites. RASMOL users with difficult or slow Internet access can download any structure from WWW-Entrez as a small PDB-format file containing only alpha-carbon or phosphorus backbone data. Alternatively, users can specify fewer models from nuclear magnetic resonance (NMR) structures containing multiple models. WWW-Entrez can also provide structures with non-degenerate coordinates, where each atom is assigned to a single location rather than an ensemble of alternatives. Such structures are useful starting points for the creation of homology models and for doing many kinds of database computations for molecular modeling.

Kinemages. WWW-Entrez has provided the first service on the WWW that can make a variety of reliable Kinemages for any structure from the public structure data. Within MAGE, clusters of buttons appear on the right of the structure image, which are used to turn parts of the structure on or off; these are scripted into the Kinemage file generated by WWW-Entrez. The MMDB atom name corrections mentioned earlier means that metals and ionic elements are found automatically and are drawn into Kinemage files in a spacefilling mode (Fig. 1b), highlighting their biological significance as a default view. This effect can only be duplicated in RASMOL (Fig. 1a) if the way the ions are named in the PDB file is known and the command-line interface is used.

Additional color. Kinemage files can also be generated from WWW-Entrez with secondary structure highlighted, and with both amino acid and nucleic acid residues colored according to chemical characteristics (hydrophobic, aromatic, charged, etc). MMDB tagging of both protein and nucleic acid chains allows the generation of Kinemage files of DNA-protein complexes in which nucleic acid and amino acid residues are properly distinguished from each other, which is not supported by the PDB parser in PreKin [4].

Coloring by 'thermal factor' was inspired by a similar feature in RASMOL and produces a Kinemage with each atom colored along a spectrum from red to blue according to its X-ray structure temperature factor. From MAGE, however, one can turn off all the atoms in each temperature layer in succession, akin to peeling an onion layer by layer.

Animations.. The animation capacity of MAGE [12] has been used by WWW-Entrez to create many animations from information buried in PDB files. These are generated automatically from either disordered portions of crystal structures of from NMR entries with multiple models (see Fig. 2). For the disordered parts of X-ray structures, bonding is calculated separately from the rest of the atoms, avoiding many of the problems arising from structures with atoms in multiple locations. For example, in the 1AKE structure, one phosphate together with Arg-167 of the adenylate kinase inhibitor, bis(adenosine-5')-pentaphosphate (Ap5A)[13] are in different conformations (Fig. 2a). Clicking on the 'ANIMATE' button swaps these two degenerate conformations, allowing the motion of the phosphate and Arg-167 to be visualized, which seem to form a salt bridge at their closest approach.

Every NMR structure in the database (now over 300) that contains multiple models can become a 'live' Kinemage animation when retrieved from WWW-Entrez (Fig. 2b). This figure is similar to the style often found in NMR structure articles [14], but it comes alive upon pressing the 'ANMIATE' button or the 'a' key. MAGE can cycle through up to 20 models one by one, creating the impression of the motions that the protein might be experiencing in solution. The animation can be played from any viewing position and the master control buttons can be used to switch on or off the 'Virtual' or 'Real' backbones, the 'Residues', heterogens or any named ions. In addition, any number or combination of models can be overlaid.

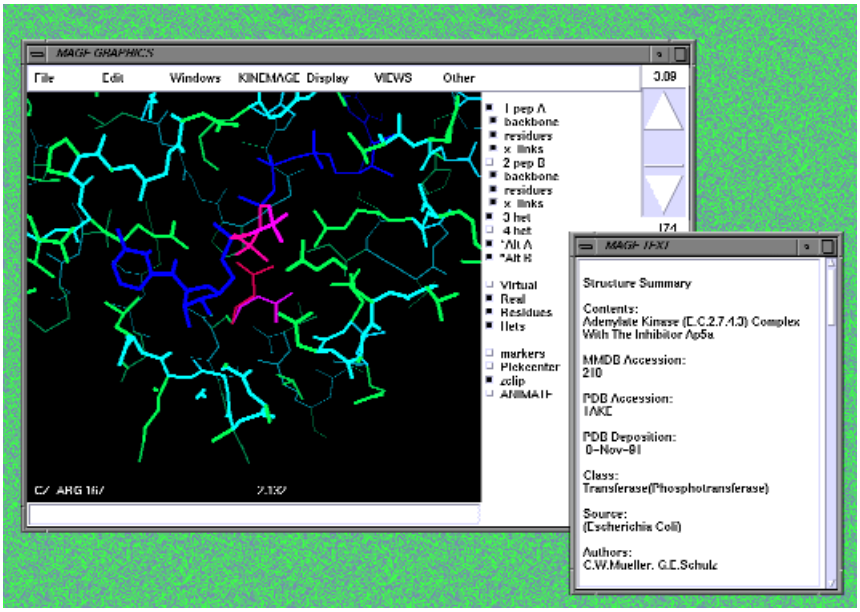

Figure 2a. A Kinemage of adenylate kinase generated using the defaults from the structure file 1AKE [13]. One of the active sites of adenylate kinase is magnified with the bound inhibitor bis(adenosine-5')-pentaphosphate (Ap5A) shown in dark blue. The magenta and red colored bonds are the two conformational-disorder ensembles containing one phosphate moiety of Ap5A and the Arg-167 residue. The ANIMATE button on the list at the right allows the user to swap the red and magenta portions illustrating the possible sidechain and inhibitor movement, which seem to form a salt bridge.

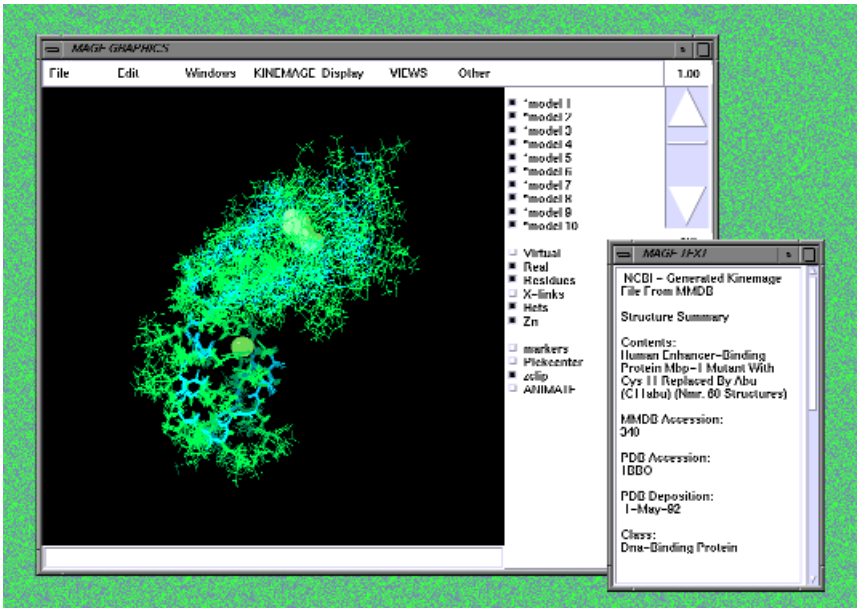

Figure 2b. A typical Kinemage of a nuclear magnetic resonance (NMR) animation obtained from the structure

1BBO [14] (human enhancer-binding protein). The Zn(2+) ions are shown, one of which is 'smeared' at the top owing to the conformational flexibility in this region. In this file, the first ten models were selected from the World Wide Web (WWW)-Entrez for incorporation into the Kinemage. Note that this figure shows the superposition of all ten models (turned on from the buttons at the right). The may be animated one-by-one using the ANIMATE button. A more compact animated file can be obtained by requesting from WWW-Entrez that only the Virtual backbone and ions from five models be put into the generated Kinemage.

The active imagery of an animation can leave the viewer with a distinct impression that proteins and biomolecules are very dynamic entities in solution, and it helps to highlight the significant contributions that NMR structures continue to make towards our understanding of structural biochemistry. For this and other reasons, we are currently completing and integrating a new three-dimensional structure viewer called Cn3d into the Network Entrez client software. This software can be used with MIME-types, to launch from a WWW-browser like RASMOL or Mage, but it will read ASN.1 data files instead of PDB- or Kinemage- format files. It will function as a direct Internet client, linking directly to NCBI servers. We will announce the availability of this viewer on the NCBI WWW site (Box 1) and describe its features at a later data.

Note - Cn3d is now available.

Box 2. Tips for viewing/retreiving structures

General

- Use the Accession field from World-Wide Web (WWW)-Entrez to retrieve by Protein Databank (PDB) ID (E.G. 1BBO).
- The 'Neighbors' button th the WWW-Entrez Structure Summary page retrieves precomputed homologous structures based on sequence, (and soon neighbors based on structure similarity.
- Once you set up RASMOL or MAGE to launch from your WWW-browser, you can use any three-dimensional service listed in Box 1.

RASMOL

- For small and fast access, select the 'Virtual Bond Model' from the WWW-Entrez Structure Summary page, these are suitable for the backbone view from RASMOL, and contain all heterogens as well.
- Select the 'All Atom Model' for a PDB file free of conformational disorder.
- Select 'Up to 5 Models' or more for the original PDB model data.
- See the RASMOL home page for custom scripts and user support (Box 1).

MAGE

- Use the WWW-Entrez Kinemage options to turn off substructures to make smaller files that transmit faster.
- Turning off 'Residues' offers substantially smaller files.
- To highlight secondary structure and residue-type, use 'Color atoms by Secondary Structure'
- Microsoft Windows - when MAGE version 3.3 (but not later versions) starts up you may hit the 'atom limit' button and increase from 3,500 to 10,000-15,000 if you have eight or more megabytes of memory.

- Macintosh - Use the System Finder. Select the MAGE Icon, then pick the 'Get Info' option from the 'File' menu. There will be a box with memory settings that you may increase.
- Unix and Linux versions of MAGE are now available (Box 1).

Acknowledgements

We thank T. Madej, F. Ouellette, and M. Boguski for helpful comments. Almost all the staff of the National Center for Biotechnology Information are in one way or another involved in the production and maintenance of the Entrez system, and although there is insufficient space to list them, we thank them all.

References

1. Richardson, D.C., and Richardson, J.S. (1992) Protein Sci. 1, 3-9 .
2. Richardson, D.C., and Richardson, J.S. (1994) *Trends Biochem. Sci.* 19, 135-138.
3. Sayle, R.A., and Milner-White, E.J. (1995) *Trends Biochem. Sci. 20*, 374-376.
4. Sokolik, C.W. (1995) *Trends Biochem. Sci.* 20, 122-124.
5. Schuler, G.D., Epstein, J.A., Ohkawa, H. and Kans J.A. (1996) *Methods Enzymol.* 266, 141-162.
6. Rose, M.T. (1990) *The Open Book, A Practical Perspective on OSI. pp. 227-322, Prentice-Hall.* .
7. Frank, J. (1996) *Three-Dimensional Electron Microscopy of Macromolecular Assembiles, Academic Press* .
8. Bryant, S.H. (1989) Protein Struct. Funct. Genet. 5 228-247 .
9. Madej, T., Gibrat, J-F., and Bryant, S.H. (1995) *Protein Struct. Funct. Genet.* 23 356-369.
10. Blancuzzi, Y., Padilla, A., Parello, J., and Cave, A. (1993) Biochemistry 32, 1302-1309. .
11. Altshul, S.F., *et al.* (1990) J. Mol. Biol, 215, 403-410. .
12. Sanchez-Ferrer, A., Nunez-Delicado, E., and Bru, R. (1995) *Trends Biochem. Sci.* 20, 286-288.
13. Muller, C.W., and Schulz, G.E. (1992) J. Mol. Biol. 224, 159-177 .
14. Omichinski, J.G., *et al*, (1992) Biochemistry 31, 3907-3917 .

NCBI Structure Cn3D

Christopher W.V. Hogue, Hitomi Ohkawa, and Stephen H. Bryant
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, Maryland U.S.A.

Rev. 17 Feb 1997
