Visualizing Lentiviral Protein 3D Structures

Brian T. Foley
Theoretical Biology and Biophysica, T10, MSK710, Los Alamos National Laboratory
Los Alamos, NM 87545 USA

Introduction

The HIV Databases at LANL have primarily focussed on primate lentiviral amino acid and genetic primary sequence information. There is increasing interest in the secondary and tertiary structures of lentiviral proteins as more 3D structures become available. The tools for viewing and manipulating these structures are also becoming more powerful and easier to use. Discovery of the 3D structures of proteins by X-ray crystallography and/or NMR is not easy. Many proteins are resistant to forming crystals, and NMR technology is limited to molecules with fewer atoms than the typical complete protein. Although there is no guarantee that a protein crystal is in the conformation that is biologically active in solution, many or most of the crystal structures in the protein structure database appear to be correct, given that data on biological activity of the proteins agrees with the structure. For one example, amino acids known through mutation analyses to influence the AZT resistance of the HIV-1 reverse transcriptase surround the AZT-binding site in the crystal structure.

The number of crystal structures for HIV-1 proteins has grown sufficiently large that it is now difficult for the inexperienced user of the PDB database to find which of the many structures is most relevant to the user. The number of software tools available for displaying 3D structures on computers is also growing, and it is difficult for potential users to decide which tool is best suited to their needs. This paper and the corresponding web pages [A] are designed to serve as an overview and guide to this growing body of data and tools.

The Databases

The Protein Data Bank (PDB) [B] is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the National Institute of Standards and Technology — three members of the Research Collaboratory for Structural Bioinformatics (RCSB)[1]. Entries in the PDB have four-digit alphanumeric entry names (such as 3HVT) which are used to access them. Each PDB entry loads into the user's web browser with the summary information page displayed. The summary page includes links to the publications which describe the structure, options for downloading and/or displaying the structure, a link to the primary amino acid sequences of the peptide chains, and several other links. The PDB home page (users should bookmark the URL of the mirror site closest to them, identified at the RSCB URL given above) provides a tutorial on searching, as well as links to search tools such as Search Lite [C], but these searches are limited by the quality of annotations provided by 3D structure authors and it is not easy to find all structures for a given protein. The tables provided herein are thus useful as a guide to HIV protein and structure entries.

The HIV-1 Protease Structures Database [D] at the National Institute of Standards and Technology (NIST) contains many protease structures which have not been submitted to the PDB. The structures from this database can be viewed as Kinemages, but not other formats such as PDB [2]. A major advantage of a protease-specific database over the PDB, is additional data fields and a search form [E], making it possible to quickly select all structures related to a single drug or strain of virus, for example.

The World Index of Molecular Visualization Resources [F] is a collection of several user-maintained databases. There are currently 13 separate databases including a database of free software, a database of commercial software, a database of image galleries and a database of biochemistry tutorials in CHIME.

The Structural Classification of Proteins (SCOP) database [G] organizes protein structures based on structural similarity [3-5]. It can be used for example to find other viral or non-viral proteases with structures similar to the HIV-1 protease. The ProSite [H] [6, 7], PFAM, [I] [8] and PRODOM [J] [9] databases are similarly useful for obtaining related groups of proteins.

A recent review of HIV-specific resources, only some of which are related to structural biology of HIV, [K] was published in 2000 [10]. The explosion in the number of databases makes any such listing or database of databases at risk of rapidly becoming obsolete. Hence the idea of user-maintained databases was implemented by Eric Martz and Trevor D. Kramer for the World Index of Molecular Visualization Resources.

The Data Structures

There are over a dozen popular formats for atomic coordinate files. One of the most popular for macromolecules is the original one used by the Protein Data Bank, commonly called a “PDB file”. The Protein Data Bank has recently adopted a new standard format, the macromolecular crystallographic information format, or mmCIF, but some popular software remains unable to process this newer format. Therefore the PDB format continues to be supported at the Protein Data Bank as well as by most other sources of atomic coordinate files. When atomic coordinate files are transmitted through the internet, their formats are identified with MIME types which have filename “extensions” such as “.pdb” to identify the type. If the user accesses a structure file with a www browser equipped with a structure viewing tool (such as the Protein Explorer, see below) installed, the PDB coordinate file can load directly into the tool and be displayed in 3D, similar to the way a “.pdf” file can be displayed by Adobe Acrobat tools if they are installed. Most of these www browser plugins allow the user to rotate, zoom in or out, and perform other manipulations of the molecule using their mouse.

Each PDB file contains a header with annotations, as well as information about the amino acid sequence of the peptides, non-peptide compounds included (such a zidovudine, glycosylated or modified amino acids, or metal ions), and a listing of the X, Y and Z coordinates for each atom represented in the structure. The mmCIF format [L] stands for macromolecular Crystallographic Information File, which allows more extensive annotation of structures, such as grouping atoms or amino acids which make up structures of biological importance. The National Center for Biotechnology Information supports an ASN.1 format [M] for structural data, which is required for the Cn3D structure viewer. Numerous other formats are in use, and each offers some specific advantages over others, at least for specific tasks. BABEL [N] is a program designed to interconvert a number of file formats currently used in molecular modeling. The program is available for Unix (AIX, Ultrix, Sun-OS, Convex, SGI, Cray, Linux), MS-DOS, and on Macs running at least System 7.0.

Viewing 3D Structures:

Software for viewing and manipulating 3D structures of proteins range from expensive and/or computationally intensive packages with capabilities such as calculating atomic forces between all atoms in a structure, to free software which “plugs in” to a www browser such as Netscape or Windows Internet Explorer. While extensive capabilities are needed to predict which chemicals in a pharmaceutical library are likely to bind to and inactivate a protein, most users do not require such computationally expensive tools. For simply viewing 3D structures, the most commonly used tool is RasMol [O]. The RasMol software has been incorporated into several other tools which add user friendly interfaces and increased functionality not found in the basic RasMol “engine”. RasMol version 2.6 is fully open-source and free [P] and RasMol version 2.7 is a continuation of this software with slightly modified restrictions on use [Q]. The Protein Explorer [R] is recommended for MacIntosh and Windows PC users. The Protein Explorer is a derivative of RasMol that uses the MDL Chime plugin [S,T]. The Protein Explorer site contains extensive on-line help for downloading and installing the Chime and Protein Explorer software. This site and other sites such as the Online Macromolecular Museum [U] also provide extensive tutorials on the capabilities and uses of this software.

Once you have installed RasMol or Protein Explorer (plus the Chime plugin), the first site a user should visit for an overview of HIV-1 structures, is the “Hall of Virology” at the Online Macromolecular Museum [V]. These exhibits include tutorials on the structure and function of HIV-1 protease, reverse transcriptase, integrase and Nef proteins. The exhibits give clear examples of how the RasMol and Chime software can be used to provide annotations of important sites and features in proteins. The Methods section at the OMM is particularly useful for its introduction to Chime scripting [W] and tutorials on using it to create museum-like displays [X]. Without these examples, loading a PDB file into RasMol or Protein Explorer for the first time far less useful, because the proteins load in single color, with no annotations of critical sites.

A more sophisticated viewer, with many more options for display, is the Visual Molecular Dynamics (VMD) software [Y]. VMD uses a different scripting language than RasMol, making it difficult to save annotated displays of proteins in one viewer for re-display in the other viewer.

A comprehensive listing of 3D molecular visualization resources is stored in a visitor-maintained database at the San Diego Supercomputing Center [Z]. Several textbooks on macromolecular 3D structure have been published in recent years [AA] .

RasMol/Protein Explorer/CHIME Scripting.

The history of macromolecular structure 3D viewing tools, and explanations of how these tools have evolved is presented in detail at [AB]. The site also provides descriptions of the advantages of each tool. An interactive tutorial on CHIME and RasMol is also available [AC]. Using a scripting language such as the RasMol/CHIME language, one can generate a specific view of any macromolecular structure, and save the script for future use. The CHIME plugin contains a “copy chime script” under the right mouse button “edit” menu item, which can be used to save the current view of a molecule, but the script that is copied with this method is very inefficient because it is “atomized” to specify the state of each atom in the molecule, rather than specifying the state of large groups of atoms that share the same state (color, spacefill, etc.). There are several mailing lists for topics related to 3D macromolecular structure scripting, such as [AD] and [AE] which are useful for learning to use structural databases and scripting languages.

Exploring Conservation of Sites.

Because of the extensive variability of existing strains and the rapid evolution of lentiviruses, tools which analyze the conservation of specific sites in proteins are of great use to HIV and SIV research. The conservation in a protein can be estimated by many measures, such as the entropy, or rates of evolution for each site. A tool to map a measurement of the site-specific rate of evolution onto each amino acid in a structure and provide various output formats including a Protein Explorer view, is the ConSurf server [AF]. The ConSurf server provides tools to score the site-specific evolutionary conservation within a protein family, and the means to map the scores onto the 3D-structure of a member of this family. The server also provides various intermediate results used in the calculations, such as the multiple alignment of the homologous sequences and the phylogenetic tree describing their relations. If the user does not provide a multiple sequence alignment, the server uses Psi-BLAST to search the SwissProt database and builds an automatic alignment. The ConSurf server translates the rate values into a 9 colors scale from 9 (highly conserved sites; colored Bordeaux) to 1 (highly variable sites; colored cyan). It then provides a Protein Explorer view of this file with the amino acids color-coded by those values and buttons than can be pressed to highlight the amino acids with each score in space-filled mode. The same method can be used to map other measurements, such as antigenicity (as measured by immunoglobulin or CTL epitope density, for example) or hydrophobicity onto the structures of HIV or SIV protein structures.

The Psi-BLAST alignment is not recommended for HIV or SIV sequences, because there are many thousands of HIV and SIV sequences in SwissProt. The Psi-BLAST method implemented by the ConSurf server will select only the few most similar sequences, and will not produce an alignment from a representative diversity of sequences. To obtain a more representative sample of HIV or SIV sequences, alignments from the HIV sequence database or alignments made by the user can be input.

Links

[A] Structures on the menu bar at http://hiv-web.lanl.gov
[B] http://www.rcsb.org/pdb/index.html
[C] http://www.rcsb.org/pdb/searchlite.html
[D] http://srdata.nist.gov/hivdb/
[E] http://srdata.nist.gov/hivdb/hivdb1.asp
[F] http://www.molvisindex.org/
[G] http://scop.mrc-lmb.cam.ac.uk/scop/
[H] http://us.expasy.org/prosite/
[I] http://www.sanger.ac.uk/Software/Pfam/
[J] http://prodes.toulouse.inra.fr/prodom/2002.1/html/home.php
[K] http://www.journals.uchicago.edu/CID/journal/issues/v31n2/000396/000396.html
[L] http://ndbserver.rutgers.edu/NDB/mmcif/index.html
[M]http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=96038927&form=6&db=m&Dopt=r
[N] http://www.ccl.net/cca/software/UNIX/babel/index.shtml
[O] http://www.umass.edu/microbio/rasmol/index.html
[P] http://www.umass.edu/microbio/rasmol/faq_ras.htm
[Q] http://www.openrasmol.org/
[R] http://molvis.sdsc.edu/protexpl/frntdoor.htm
[S] http://www.mdl.com/chime/index.html
[T] http://www.umass.edu/microbio/chime/index.html#about
[U] http://www.clunet.edu/BioDev/omm/gallery.htm
[V] http://www.clunet.edu/BioDev/omm/exhibits.htm#displays
[W] http://www.clunet.edu/BioDev/omm/scripting/molmast.htm
[X] http://www.clunet.edu/BioDev/omm/howdo.htm
[Y] http://www.ks.uiuc.edu/Research/vmd/
[Z] http://molvis.sdsc.edu/visres/index.html
[AA] http://molvis.sdsc.edu/protexpl/favlit.htm#books
[AB] http://www.umass.edu/microbio/rasmol/history.htm
[AC] http://www.chem.uwec.edu/ChimeTutDemos/tuts/rasmol.html
[AD] http://www.umass.edu/microbio/rasmol/raslist.htm
[AE] http://www.ks.uiuc.edu/Research/vmd/mailing_list/
[AF] http://consurf.tau.ac.il/

Index to HIV Macromolecular Structures

3D STRUCTURES On the Menu Bar at http://hiv-web.lanl.gov

Tutorials:

The Online Macromolecular Museum http://www.clunet.edu/BioDev/omm/exhibits.htm Provides turorials on several HIV proteins.

Teresa Larsen: An Accurate Look into HIV http://www.sdsc.edu/GatherScatter/GSsummer96/larsen.html demonstrates use of structure information.

Selected pictures of HIV-1 reverse transcriptase (RT) by Kalyan Das http://www.cabm.rutgers.edu/~kalyan/RT_imgs/index.html provides images of HIV-1 RT structures.

The PDB highlighted Reverse Transcriptase as molecule of the month for Sept 2002 http://www.rcsb.org/pdb/molecules/pdb33_1.html

Review Articles:

Turner BG, Summers MF. Structural biology of HIV. *J Mol Biol* 1999 Jan 8;**285**(1):1–32 PMID: 9878383

Huang H, Chopra R, Verdine GL, Harrison SC. Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science*, 1998 Nov 27;**282**(5394):1669–75. PMID: 9831551

Tools:

World Index of Molecular Visualization Resources http://www.molvisindex.org is a user-maintained database of tools and resources.

The ConSurf and Server http://consurf.tau.ac.il provides tools to score the site-specific evolutionary conservation within a protein family, and the means to map the scores onto the 3D-structure of a member of this family.

The ConSeq server http://conseq.bioinfo.tau.ac.il is useful for the identification of functionally and structurally important residues in protein sequence alignments.

Protein Explorer (PE) http://www.proteinexplorer.org which uses a web browser and the free CHIME plugin

http://www.mdlchime.com to render 3D images of protein structures from the PDB database http://www.rcsb.org/pdb is recommended as one of the best tools for exploring macromolecular structures. A description of the benefits of PE can be found on the PE site http://molvis.sdsc.edu/protexpl/why_pe.htm .

References

1. Berman, H.M., et al., The Protein Data Bank. *Nucleic Acids Research*, 2000. **28**:235–242.
2. Vondrasek, J. and A. Wlodawer, HIVdb: a database of the structures of human immunodeficiency virus protease. *Proteins*, 2002. **49**(4):429–31.
3. Murzin, A.G., et al., SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 1995. **247**(4):536–40.
4. Lo Conte, L., et al., SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res*, 2002. **30**(1):264–7.
5. Lo Conte, L., et al., SCOP: a structural classification of proteins database. *Nucleic Acids Res*, 2000. **28**(1):257–9.
6. Sigrist, C.J., et al., PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*, 2002. **3**(3):265–74.