SHORT COMMUNICATION

Application of InChI to Curate, Index, and Query 3-D Structures

M.D. Prasanna,[1] Jiri Vondrasek,[2] Alexander Wlodawer,[3] and T.N. Bhat[1]*
[1]*Biotechnology Division (831), NIST, Gaithersburg, Maryland*
[2]*Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Prague, Czech Republic*
[3]*Macromolecular Crystallography Laboratory, National Cancer Institute, Frederick, Maryland*

ABSTRACT **The HIV structural database (HIVSDB) is a comprehensive collection of the structures of HIV protease, both of unliganded enzyme and of its inhibitor complexes. It contains abstracts and crystallographic data such as inhibitor and protein coordinates for 248 data sets, of which only 141 are from the Protein Data Bank (PDB). Efficient annotation, indexing, and querying of the inhibitor data is crucial for their effective use for technological and industrial applications. The application of IUPAC International Chemical Identifier (InChI) to index, curate, and query inhibitor structures HIVSDB is described. Proteins 2005;60:1–4.**
Published 2005 Wiley-Liss, Inc.*

INTRODUCTION

The cure for AIDS is still far from a reality, since almost all current methods of treatment fall into the category of either containment or prevention. Advancing our knowledge of the human immunodeficiency virus (HIV) is a necessary step in the development of effective treatments of this disease. Currently, much of the research on the treatment of AIDS is directed either toward vaccine development or toward drug development. Although several promising leads on vaccine have been reported,[1,2] no effective vaccine has been developed so far. Another approach for AIDS treatment is by the use of drugs that selectively inhibit virally encoded enzymes, such as reverse transcriptase or HIV protease, the latter a member of the aspartic protease family.[3] In fact, such drugs, together with fusion inhibitors, provide the only clinically proven method for the treatment of AIDS. Drugs specifically designed to inhibit HIV protease, an enzyme that carries out the necessary post-translational processing of the viral gag-pol polypetide into functional viral components, have been very successful. The processing of the gag-pol translational product by HIV protease releases the viral replication enzymes (protease, reverse transcriptase/ribonuclease H, and integrase).[4] This activity is essential for the viral life cycle, and therefore disrupting the proteolytic activity through the application of inhibitors results in noninfectious virions.[5] For this reason, a concentrated effort of many laboratories has gone into the elucidation of enzyme/inhibitor interactions of this enzyme (as well as of the other HIV enzymes), and extensive efforts have focused on developing strategies for disrupting critical macromolecular interactions required for the viral life cycle.[6] Thus, a critical need for structural information on these systems will exist as long as drug development for the treatment of AIDS represents work in progress. Informatics and infrastructure support for such activities are key to the success of this work. Several databases dealing with AIDS-related issues have been established. The National Institute of Allergy and Infectious Diseases (NIAID), a unit of the National Institutes of Health, has created a searchable database of chemical and biological anti-HIV compounds that is freely accessible to all researchers (http://www.niaid.nih.gov/daids/dtpdb/intro.htm). The Los Alamos National Laboratory has developed a web resource of biological sequences, drug resistance mutations and vaccine trials (http://www.hiv.lanl.gov/content/index). The curated data resource (http://hivdb.stanford.edu/) focuses on HIV drug resistance. While these efforts, as well as some others, provide very valuable information, they do not concentrate on any particular class of drug design targets.

For this reason, we have developed HIVSDB, a structural resource that includes comprehensive data derived from the structures of HIV protease and, particularly, of the inhibitor complexes of this enzyme. The resource was originally established at the National Cancer Institute (NCI), Frederick, MD, using static HTML web pages.[7] The database contains extensive data on the enzyme and its inhibitor complexes, much of it not available elsewhere, and is thus a unique resource in the continuing efforts of design and development of novel drugs that would be

Availability: http://xpdb.nist.gov/hivsdb/hivsdb.html

*Correspondence to: T.N. Bhat, Biotechnology Division (831), NIST, 100 Bureau Drive, Gaithersburg, MD 20899-8314. E-mail: bhat@nist.gov

Received 6 August 2004; Accepted 22 November 2004

Published online 28 April 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20469

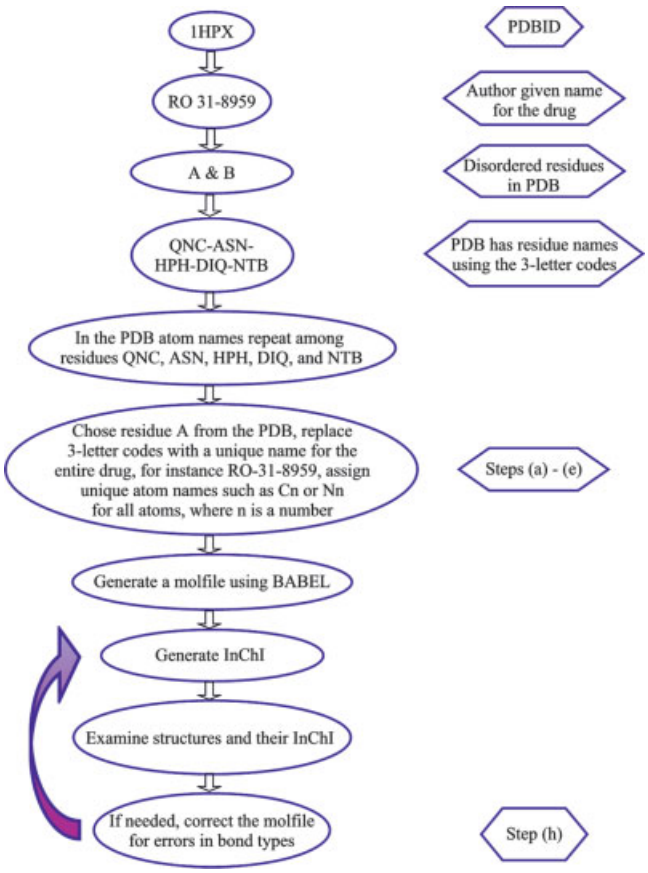


Fig. 1. An example of the steps used to assign InChI for an inhibitor RO-31-8959 (PDBID 1HPX).

useful to control the AIDS epidemic. Extensive efforts to develop novel chemical compounds have been going on both in academia and in industry for close to 20 years and they have resulted in a multitude of drug candidates (see http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/index.html). However, navigating through these lists of compounds is difficult, since most of them have not been labeled using uniform rules. Analysis of the database of compounds and comparison of their activities against HIV protease requires a versatile technique to index, compare these inhibitors. This paper describes and illustrates a novel technique that was developed for this purpose.

The emergence of web-based information-handling resources has had an enormous impact on data exchange. Although the ease with which inhibitor data could be exchanged among researchers is phenomenal, the ability to correlate data using names of inhibitors is still a problem.[8] The IUPAC International Chemical Identifier[9] (InChI, http://www.iupac.org/projects/2000/2000-025-1-800.html) promises a solution to this problem.

InChI is an automatically created unique index for a chemical compound. This index is generated using chemical properties such as atom types, nature of bonds, chirality, and atomic charge of the molecule. This index is stored using text strings that are organized in several "layers" corresponding to different varieties of structural information. This layered arrangement of InChI not only allows the software to gradually build the chemical identifiers in a series of well-defined steps, but also allows a user to selectively utilize these layers for data annotation and navigation purposes. For certain compounds, the available data may not be accurate in all layers. For instance, bond lengths derived from low-resolution X-ray data may not be

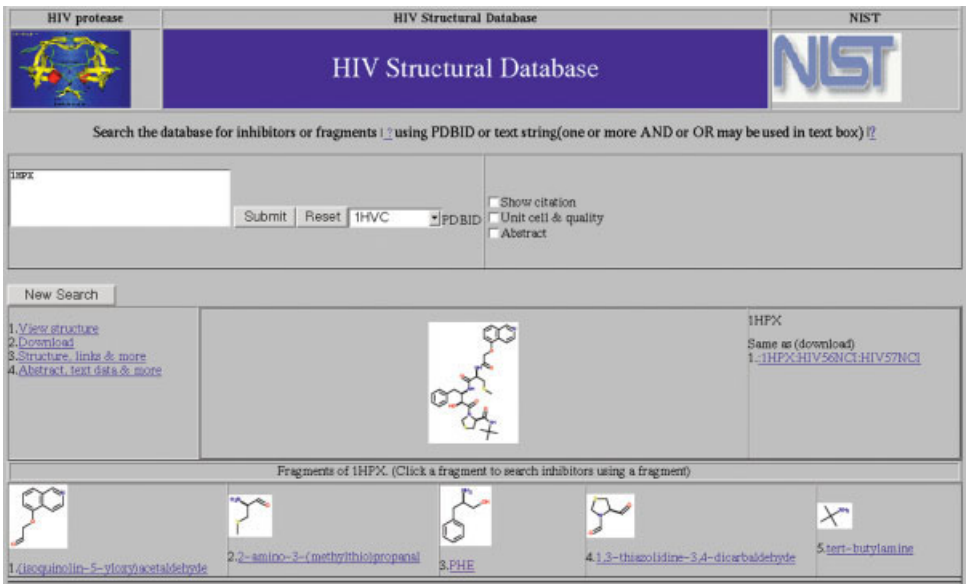
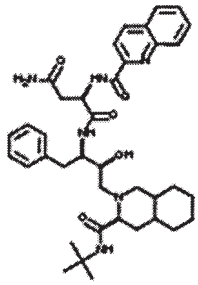


Fig. 2. A query for the structures corresponding to PDB accession code (PDBID) 1HXB. The web tool shows the inhibitor molecule resulting from the query at the center of the page and lists all entry IDs (1HXB, HIV23NCI, HIV82NCI) that contain an inhibitor identical to that of the query (1HXB). These identical inhibitors are identified and indexed using InChI. Users may download data or view structures for these entries using the hyperlinks.

TABLE I. InChI and a Molecular Sketch of Saquinavir, the First Inhibitor of HIV Protease to be Approved as an AIDS Drug†

PDBID	InChI	Molecule
1HXB	C38H50N6O5,1H3-38(2H3,3H3)43H-31(48)35H-21H2-32H-17H2-15H2-16H2-18H2-33H(32)22H2-44(35)23H2-37H(49H)36H(19H2-24-9H-5H-4H-6H-10H-24)42H-30(47)34H(20H2-28(39H2)45)41H-29(46)27-14H-13H-25-11H-7H-8H-12H-26(25)40-27 32−,33+,34−,35−,36−,37+	

†InChI is shown using XML.

accurate enough to determine their bond types uniquely. In such situations, one may compare InChI of other layers among related compounds to detect and correct the errors in bond types. When used in this way, InChI may also provide a validating and annotating capability for inhibitors and ligands found in macromolecular structures. In other situations, uncertainty may exist in a stereochemical center—such uncertainty can be expressed in a consistent, robust manner. To test and to illustrate the use of InChI to macromolecular structural data, we chose to apply it to the inhibitors found in HIVSDB.

RESULTS AND DISCUSSION

The application of InChI to the inhibitors of HIVSDB was carried out as follows (Fig. 1). (1) The HIVSDB database holds the atomic coordinates in the Protein Data Bank (PDB) format.[10] From this database extract the 3-D coordinates and the atom names of all atoms of the protease inhibitors. (2) Identify and remove the disordered atoms from this list of coordinates and atom names. (3) Examine all the atom names to ensure that they always start with atomic symbols. This convention is not often followed by the PDB. For instance, the PDB often wraps around letters of atom names to accommodate names that require additional characters to identify them uniquely within a residue. (4) Sometimes, X-ray crystallographic studies require the inclusion of two or more molecules of the same inhibitor bound to two or more chemically identical but structurally different protein molecules. However, the current InChI version does not distinguish between 3-D structural features such as molecular conformations. Therefore, identify and remove multiple occurrences of an inhibitor in a given entry. (5) In a convention used sometimes (and inconsistently) by the PDB, inhibitors are annotated[11,12] into their fragments, and assigned individual residue numbers for each of these fragments. In order to work correctly, the software that generates InChI requires that all fragments of a compound be collated to form a single inhibitor molecule. Examine and collate all fragments of each inhibitor molecule. If necessary, rename the atoms of individual fragments to avoid duplicate atom names within an inhibitor molecule. (6) Generate Molfile (http://www.mdl.com/downloads/public/ctfile) for each inhibitor molecule using BABEL (http://openbabel.sourceforge.net/). (7) For each molfile assign an InChI. (8) The 3-D coordinates result from refinement of crystallographic data with programs such as X-PLOR[13] that regularize bonds using restraints on bond distances. Quite often, primarily due to lack of high resolution data, such refinements produce bond distances that are not sufficient for BABEL to assign their bond types correctly. These irregularities may result in errors (wrong assignment of single, double, or partial bond attributes) in molfiles, and thus errors in the assignment of InChI–identical inhibitors with nonidentical InChI. Annotate such errors by first comparing the InChI at "reduced resolution" using layers such as those for just the chemical formula of the compounds. Whenever an error in the assignment of InChI is detected, edit the molfile using CHEMDRAW[14] and assign a new InChI (Table I).

InChI, together with other information such as the abstract, unit-cell parameter, space group, IUPAC name of an inhibitor and entry id (PDBID) were loaded into an ORACLE table and a web interface was developed using PERL and Structured Query Language (SQL). This web interface provides several different applications of the InChI to link up identical inhibitors from different entries that result from a query (Fig. 2).

In this era of Internet and web-based technologies, interoperability and federated approaches are necessary evils. Exchange of data on chemical compounds is complicated due to their nature and complexity. Attempts to unify names of chemical compound using IUPAC rules have met with only partial success.[15] Recent progress and release of InChI by IUPAC and NIST as an indexing standard for chemical compounds based solely on chemical structural data opened up a novel, unique opportunity for enhancing interoperability of chemical data. Exchange and comparison of data on ligands and inhibitors in macromolecular structures has been impaired due to the lack of such a reliable indexing procedure and here we illustrate the application of the InChI for such purposes. Such tools will be particularly useful to direct design and synthesis of novel compounds with improved pharmacolog-

ical properties—needed due to the rapid development of resistance against the current generation of AIDS drugs. An example of detailed analysis that might be significantly simplified and improved with the tools described here was recently provided for HIV protease.[16] Similar approaches may also be used in the future for other drug targets if the use InChI becomes more widespread.

ACKNOWLEDGMENTS

We thank M. Tung, F. Schwarz, R. Goldberg, H. Rodrigues, G. Gilliland, D. Tchekhovskoi, S. Stein, and V. Vilker for helpful discussions and support. Partial funding of this effort was provided by the Systems Integration for Manufacturing Applications (SIMA) and also by the Exploratory Research Award for the year 2004. Certain trade and company products are identified in this paper to specify adequately the computer products needed to develop this data system. In no case does such identification imply endorsement by the National Institute of Standards and Technology (NIST), or does it imply that the products are necessarily the best available for the purpose.

REFERENCES

1. Shiver JW, Fu TM, Chen L, Casimiro DR, Davies ME, Evans RK, Zhang ZQ, Simon AJ, Trigona WL, Dubey SA, and others. Replication-incompetent adenoviral vaccine vector elicits effective anti-immunodeficiency-virus immunity. Nature 2002;415:331–335.
2. Barouch DH, Kunstman J, Kuroda MJ, Schmitz JE, Santra S, Peyerl FW, Krivulka GR, Beaudry K, Lifton MA, Gorgone DA, and others. Eventual AIDS vaccine failure in a rhesus monkey by viral escape from cytotoxic T lymphocytes. Nature 2002;415:335–339.
3. Gulick RM, Mellors JW, Havlir D, Eron JJ, Gonzalez C, McMahon D, Richman DD, Valentine FT, Jonas L, Meibohm A, and others. Treatment with indinavir, zidovudine, and lamivudine in adults with human immunodeficiency virus infection and prior antiretroviral therapy. N Engl J Med 1997;337:734–739.
4. Kohl NE, Emini EA, Schleif WA, Davis LJ, Heimbach JC, Dixon RA, Scolnick EM, Sigal IS. Active human immunodeficiency virus protease is required for viral infectivity. Proc Natl Acad Sci 1988;85:4686–4690.
5. Lambert DM, Petteway SR, Jr., McDanal CE, Hart TK, Leary JJ, Dreyer GB, Meek TD, Bugelski PJ, Bolognesi DP, Metcalf BW, and others. Human immunodeficiency virus type 1 protease inhibitors irreversibly block infectivity of purified virions from chronically infected cells. Antimicrob Agents Chemother 1992;36:982–988.
6. Turner BG, Summers MF. Structural biology of HIV. J Mol Biol 1999;285:1–32.
7. Vondrasek J, Wlodawer A. HIVdb: a database of the structures of human immunodeficiency virus protease. Proteins 2002;49:429–431.
8. Freemantle M. Unique labels for compounds. C&EN London 2002;80:33.
9. Stein SE, Tchekhovskoi D, Heller SR. The IUPAC and NIST Chemical Identifier. 2004.
10. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
11. Bhat TN, Bourne P, Feng Z, Gilliland G, Jain S, Ravichandran V, Schneider B, Schneider K, Thanki N, Weissig H, and others. The PDB data uniformity project. Nucleic Acids Res 2001;29:214–218.
12. Westbrook J, Feng Z, Jain S, Bhat TN, Thanki N, Ravichandran V, Gilliland GL, Bluhm W, Weissig H, Greer DS, and others. The Protein Data Bank: unifying the archive. Nucleic Acids Res 2002;30:245–248.
13. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, and others. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54:905–921.
14. ACDLAB. http://www.acdlabs.com/products/chem_dsn_lab/chemsketch/tech.html ACD/ChemSketch IUPAC names.
15. Adam D. Chemists synthesize a single naming. Nature 2002:369.
16. King NM P-JM, Nalivaika EA, Schiffer CA. Combating susceptibility to drug resistance; lessons from HIV-1 Protease. Chem Biol 2004;11:1333–1338.