Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

PDB Editor: a user-friendly Java-based Protein Data Bank file editor with a GUI

Jonas Lee[a,b] and Sung-Hou Kim[a,b]*

[a]Department of Chemistry, University of California, Berkeley, California 94720-5230, USA, and [b]Physical Bioscience Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

Correspondence e-mail: shkim@cchem.berkeley.edu

Received 29 September 2008
Accepted 7 February 2009

The Protein Data Bank file format is the format most widely used by protein crystallographers and biologists to disseminate and manipulate protein structures. Despite this, there are few user-friendly software packages available to efficiently edit and extract raw information from PDB files. This limitation often leads to many protein crystallographers wasting significant time manually editing PDB files. *PDB Editor*, written in Java Swing GUI, allows the user to selectively search, select, extract and edit information in parallel. Furthermore, the program is a stand-alone application written in Java which frees users from the hassles associated with platform/operating system-dependent installation and usage. *PDB Editor* can be downloaded from http://sourceforge.net/projects/pdbeditorjl/.

1. Introduction

With the exponential growth in the number of macromolecular structures in the Worldwide Protein Data Bank (Berman *et al.*, 2003) during the present structural genomics era, the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977) coordinate file format has become the computer file format most widely used by structural biologists, bioinformaticists and biologists to disseminate, manipulate and analyze macromolecular atomic coordinates. For protein crystallographers, the PDB file format is often the workhorse format that is used in almost all software suites that manipulate atomic coordinates in the structure-determination process, including the traditional *CCP*4 suite (Collaborative Computational Project, Number 4, 1994) and the newly developed *PHENIX* package (Adams *et al.*, 2002). Despite this, there are not many user-friendly software packages available to quickly and effectively edit PDB files.

Although many new crystallographic software suites such as *PHENIX* and *HKL*-3000 (Minor *et al.*, 2006) and modeling software such as *Coot* (Emsley & Cowtan, 2004) are becoming user-friendly and automated in order to require less user intervention, crystallographers are not yet completely free from manually editing their coordinate files to solve a crystal structure. Text-manipulation script languages such as Awk and Perl are available, as well as several popular programs for the modification of PDB files such as *PDBSET* (Collaborative Computational Project, Number 4, 1994) and *MOLEMAN* (Kleywegt & Jones, 1997). However, Awk and Perl require some proficiency in computer programming, which often becomes a significant barrier. Nevertheless, we have found that many crystallographers still rely on basic text editors to perform repetitive tasks including trimming unwanted atoms, changing chain IDs and residue numbers and resetting temperature factors and occupancies. These manual interventions often lead to great frustration among novice protein crystallographers because the PDB file format employs fixed-width data fields. Hard-to-debug format errors, such as the insertion or deletion of a required white-space character, can easily be introduced.

A previous effort to overcome these problems was accomplished by adding a module for editing PDB files to the EMACS text editor (Bond, 2003). This is an excellent adaptation of the highly customizable open-source EMACS text editor that provides experienced protein crystallographers who are used to editing with text editors with new tools to expedite their work. Although not a PDB editor, the mmCIF editor available in the *mmLib* toolkit (Painter & Merritt, 2004) is also an excellent GUI coordinate-file editor. Unlike these two editors, our editor tries to specialize in making the editing task easier and faster for novice crystallographers. In order to achieve this, we made our program a stand-alone Java program that displays PDB data in a spreadsheet-like tabulated format and all the available functions and commands are accessible through Swing GUI. Our editor is also specialized for sort, search, parallel edit and selective manipulations, which are useful functions that have not been implemented in other editors. The editor also supports imports and exports of data to or from spreadsheet format by clipboard actions. Data can be easily transferred to common spreadsheet programs such as *Excel* and *OpenOffice Calc*.

2. Description

PDB Editor was designed with Java Swing GUI in order to maximize its user-friendliness and portability between operating systems (Fig. 1). The current program offers all the common user-friendly features including file open/save dialog boxes, save confirmations before exiting, clipboard commands including cut, copy and paste, and undo/redo commands for data manipulation. The editor will display coordinate-related records (ATOM, HETATM, ANISOU, SIGATM, SIGUIJ) in a spreadsheet-like table format. Connectivity (CONECT) records are directly interfaced with the coordinate data and are automatically updated as atom information changes. Although they are not directly interfaced with the coordinate data, noncoordinate data such as the title, remark, heterogen, primary

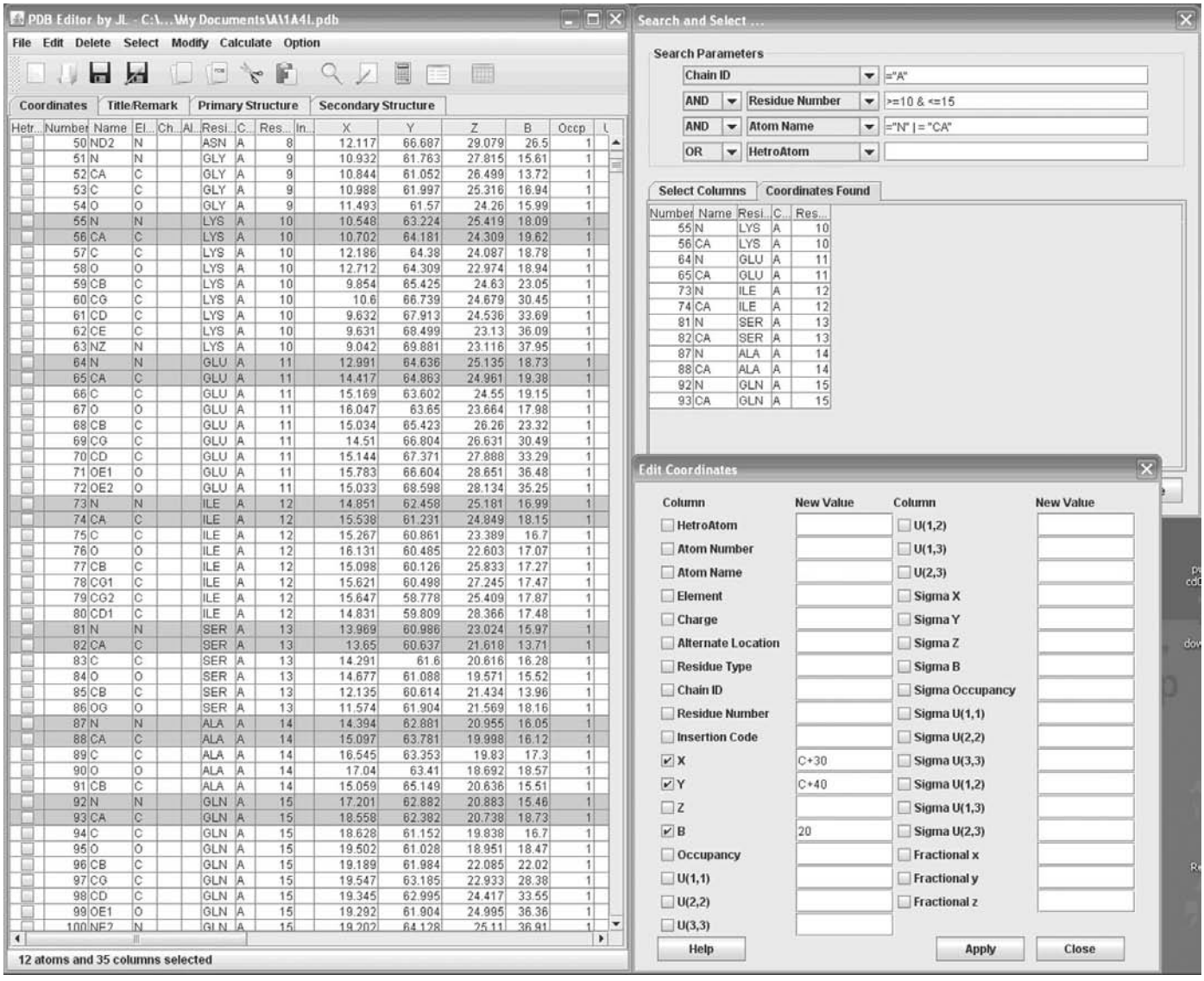

Figure 1
PDB Editor main window with Search and Select and Edit Coordinate subwindows running on Windows XP. The speed and selectivity of editing is maximized by a conditional search of atoms using a Search and Select subwindow and by parallel editing with a single fixed value, or by mathematical operations referencing data values from a particular atom in the Edit Coordinates subwindow.

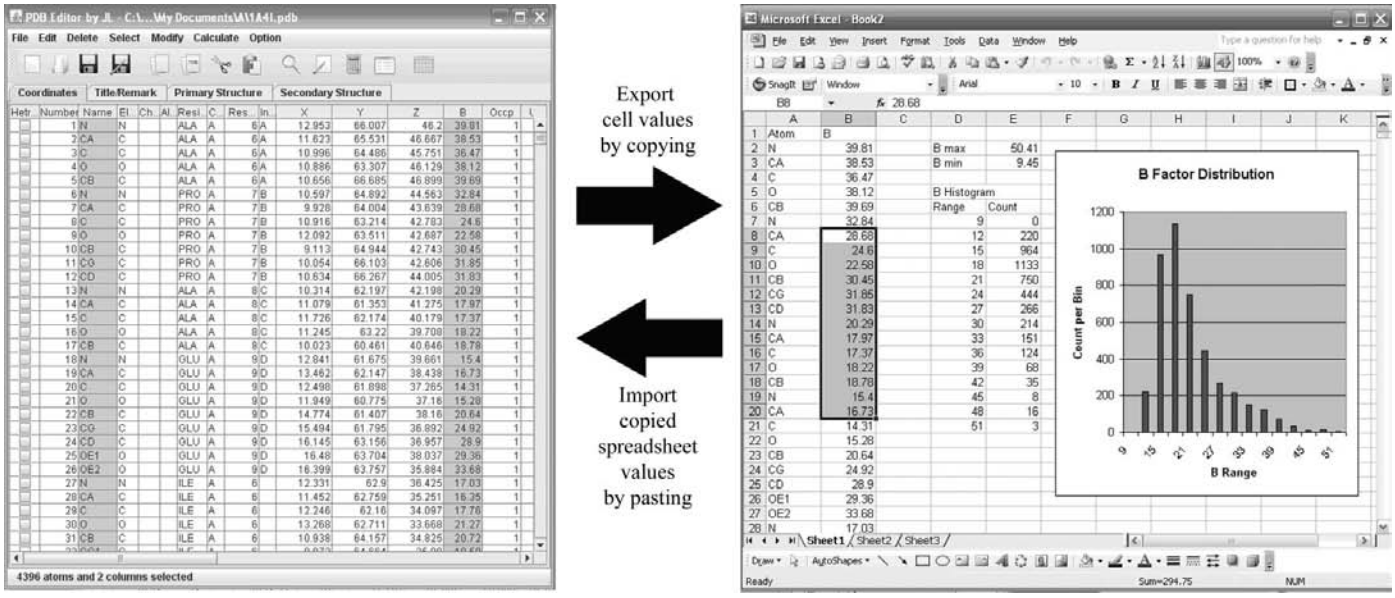

Figure 2
Data exchange between *PDB Editor* window and *Microsoft Excel* using clipboard commands. The editor's extensibility is further achieved by allowing quick data transfer between the editor and spreadsheet or database tables by the simple clipboard actions of copying and pasting. The values of the exported data can be used for a quick analysis such as that depicted in the figure. The data can also be edited in the spreadsheet and reimported back into the editor by copying and pasting.

structure and secondary structure are displayed in a table list format for editing. The current version of *PDB Editor* parses and allows the user to edit all records specified in PDB format version 2.3 (although some, such as TER, only indirectly); this excludes MODEL, ENDMDL, MTRIX and TVECT records. Multi-model PDB file support is currently under development.

All functions and commands are accessible through a GUI, which frees users from looking through a manual to search, find and type in the appropriate command. All of the basic and widely used functions and commands are linked to keyboard shortcuts to expedite data manipulation. Clipboard actions are divided into PDB-format manipulation and tab-delimited text manipulation. PDB-format clipboard manipulation can cut, copy and paste atoms between *PDB Editor* windows and other text editors, while tab-delimited text clipboard manipulations can export and import data to and from external spreadsheet programs for specialized manipulation and analysis (Fig. 2).

Aside from the basic editing of individual cells in the data table, the maximum editing capability of the editor can be achieved by using the following functions: conditional coordinate search and parallel editing and manipulation of selected atoms. These two functions are controlled by two subwindows: the Search and Select window and the Edit Coordinates window (Fig. 1). The Search and Select window supports general and advanced searches by user-defined conditions and parameters. The search conditions support common mathematical comparison operators, including >, < and =, to pick atoms by data value. Atoms selected by this method (or manually) can be edited in parallel using the Edit Coordinate window. With this window, parallel manipulations can be made using either a user-inputted value or a mathematical formula. The mathematical operation supports all the basic mathematical operations, including addition, subtraction, multiplication and division, as well as allowing referencing of data values from other atoms. Also, users can instantly switch values between two columns, which is useful for fractional coordinate manipulation.

PDB Editor has extensive features for further coordinate manipulations. These include basic analysis tools such as a sequence extractor, an atom-distance calculator and a data-statistics calculator. Other manipulation tools include a symmetry-mate generator, a quick matrix parser that can rotate/translate atoms (ANISOU is not yet supported), a residue-number and chain-ID editor with automatic residue-number conflict detector and fixer, a connectivity-record editor tool with automatic atom tracker, an atom-order correction tool with automatic order detection of residues with an insertion code by searching for closest peptide-bond connections, an atom sorter and atom-number resetter and a tool to reduce a protein to C^{α} or the alanine backbone (glycines are kept as glycines in alanine backbone reduction). Furthermore, all the functions listed above provide a hassle-free means of selective manipulation. All of the manipulations can be performed without the need for several intermediate files.

3. Program implementation

PDB Editor is a stand-alone application written in portable Java Swing GUI. This program can be run on any platform with Java SE 5 or higher, which includes Windows XP/Vista, Mac OS X 10.4 or greater and Linux. Java can be downloaded from http://www.java.com for Windows and Linux users and is updated at http://www.apple.com/java for Mac OS X users. The program, manual and source code are licensed by GNU Public License (GPL) and are freely available without any download or use restrictions from http://sourceforge.net/projects/pdbeditorjl/. Help and support are provided through SourceForge.net discussion boards, which can be accessed through the URL given above.

The authors are grateful to Drs Jose Henrique Pereira and In-Geol Choi for their thoughtful suggestions and encouragement, Dr Gerard 'DVD' Kleywegt for critical evaluation and advice, and Dr Gregory E. Sims and Barbara Gold for proofreading the manuscript.

References

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *Acta Cryst.* D**58**, 1948–1954.

Berman, H. M., Henrick, K. & Nakamura, H. (2003). *Nature Struct. Biol.* **10**, 980.
Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
Bond, C. S. (2003). *J. Appl. Cryst.* **36**, 350–351.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Kleywegt, G. J. & Jones, T. A. (1997). *Methods Enzymol.* **277**, 208–230.
Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. (2006). *Acta Cryst.* D**62**, 859–866.
Painter, J. & Merritt, E. A. (2004). *J. Appl. Cryst.* **37**, 174–178.