Bioinformatics, 31(17), 2015, 2894–2896
doi: 10.1093/bioinformatics/btv223
Advance Access Publication Date: 24 April 2015
Applications Note

Structural bioinformatics

PDBest: a user–friendly platform for manipulating and enhancing protein structures

Wellisson R. S. Gonçalves[1,*], Valdete M. Gonçalves-Almeida[1], Aleksander L. Arruda[1], Wagner Meira Jr.[1], Carlos H. da Silveira[2], Douglas E. V. Pires[3,*,†], Raquel C. de Melo-Minardi[1,*,†]

[1]Department of Computer Science, Universidade Federal de Minas Gerais, Brazil, [2]Advanced Campus at Itabira, Universidade Federal de Itajubá, Brazil and [3]Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz, Brazil

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint First Authors.

Received on March 6, 2015; revised on April 10, 2015; accepted on April 19, 2015

Abstract

Summary: PDBest (PDB Enhanced Structures Toolkit) is a user-friendly, freely available platform for acquiring, manipulating and normalizing protein structures in a high-throughput and seamless fashion. With an intuitive graphical interface it allows users with no programming background to download and manipulate their files. The platform also exports protocols, enabling users to easily share PDB searching and filtering criteria, enhancing analysis reproducibility.
Availability and implementation: PDBest installation packages are freely available for several platforms at http://www.pdbest.dcc.ufmg.br
Contact: wellisson@dcc.ufmg.br, dpires@dcc.ufmg.br, raquelcm@dcc.ufmg.br
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Acquiring, manipulating and filtering large sets of biomolecule files from repositories such as the Protein Data Bank, is a common and important preprocessing step prior to protein structure studies. By treating these files one aims to improve data quality, identifying and/or removing common idiosyncrasies that could compromise and bias subsequent analysis. Common filtering tasks include, but are not limited to: filtering multiple occupancies and models, splitting files by chain, selecting subsets of atoms or residues, identifying missing atoms or residues, generating an assessment report.

Given the relevance of Structural Biology and Bioinformatics, several initiatives have been successful in developing software tools and libraries for manipulating PDB files for different programming languages, such as Python (Cock *et al.*, 2009), Java (Holland *et al.*, 2008), R (Grant *et al.*, 2006) and others (Gajda, 2013), which are widely used today.

It is, however, a task of great relevance for both educational and scientific purposes to make the manipulation and study of such source of data more accessible to a larger public (from undergraduate students to researchers) without requiring a programming background, while taking into account the ever increasing scale of current data availability and guaranteeing the reproducibility of the process.

Despite the efforts of the community to make these capabilities available as web services (Dolinsky *et al.*, 2007; Hussain *et al.*, 2002; Wang *et al.*, 2005; Wiederstein and Sippl, 2007), there still is a significant demand for an easy-to-use, scalable platform for treating protein structures.

To fill this gap, we have developed PDBest (PDB Enhanced Structures Toolkit), a user-friendly, freely available platform for acquiring, manipulating and normalizing protein structures in a reproducible, high-throughput and seamless fashion.

2 Platform description and workflow

PDBest was developed aiming to achieve four major goals: (i) provide means to improve data quality in Structural Biology/Bioinformatics; (ii) make it accessible for a larger public; (iii)

provide scalability and parallel processing; (iv) as well as reproducibility and reusability.

To establish a high-quality structural database, PDB files of interest need to be collected and properly preprocessed and/or cleaned. Many of these cleaning tasks are common to different studies (as the example in Fig. S1 of Supplementary Material), including treating NMR models or multiple occupancies while others are more specific (e.g. filtering subsets of residues or atoms).

To cope with these demands PDBest provides an intuitive graphical user interface (GUI), as depicted on Figure 1A. It was designed to make these tasks accessible to a broader public since no command line or programming experience is require, and also taking advantage of the emergence of multi-core architectures, dramatically reducing processing time via parallel processing. The software platform was developed in C++ language on the QT framework, providing high performance for all major operating systems: Windows, Unix/Linux and Mac OS X. Many capabilities, including adding hydrogens and file format conversions, were implemented using the OpenBabel library (OLBoyle *et al.*, 2011).

To achieve reproducibility, we have developed the PDBest protocol file. The protocol file encompasses all user-defined filtering preferences, data collection and filtering options that can be easily shared, improving work reproducibility and the creation of filtering standards. Figure 1B shows PDBest capabilities and pipeline which is divided in the stages described next.

Data Acquisition: Structure files can be loaded into PDBest in three ways: (i) by submitting an 'Online Query' to the RCSB PDB mirror; (ii) by loading local files or (iii) a combination of both. Using (i) users are able to search for structures using all parameters available at the RSCB web site, performing complex queries including logical operators 'and'/'or' in any combination. No restriction on the number of PDB files to be acquired is imposed. Online queries are also compatible with those of well established servers, like the PISCES web server (Wang *et al.*, 2005). Figure 1A shows how a PISCES-like query can be performed using PDBest.

Filtering Standards: After acquiring the PDB files users can set preprocessing steps to be applied. It is possible to edit or filter any information described in the PDB File Format Documentation, as well as converting between known biological file formats. Amongst the main filtering capabilities are included: splitting files by chain, selecting models, parsing ligands, adding/removing hydrogens, treating multiple occupancies, renumbering residues and atoms as well as selecting subsets of atoms or residues of interest.

Output and Processing: Processed PDB files are generated, with new suffixes defined by the user. A detailed report can be presented highlighting files with identifies issues (missing residues, missing atoms, multiple occupancies or any inconsistencies).

Reproducibility with PDBest—Protocol file: PDBest can also generate a protocol file which stores user section information, filtering parameters, online query and steps employed to generate the data set. This enables users to reuse and share data acquisition and preparation steps, improving reproducibility of the work and subsidizing the creation of processing standards. A tutorial with examples of usage is available as Supplementary Material.

3 Conclusion

PDBest offers a powerful, user-friendly interface between researchers/students in Structural Biology and common processing tasks for protein structural files. Its graphical interface provides an efficient, yet accessible, way to acquire, manipulate and check PDB files as means to improve data quality and integrity.

The platform is capable of parallel processing, taking advantage of current multi-core architectures and providing scalability for its pipeline. On top of that, PDBest enables users to share acquisition and filtering protocols, improving work reproducibility and the creation of filtering standards for research groups.

PDBest is available for all major Operating Systems and has been thoroughly tested and proved fundamental on many structure-based studies in our group including: inter-residue contact analysis (da Silveira *et al.*, 2009), receptor-based ligand prediction (Pires *et al.*, 2013),

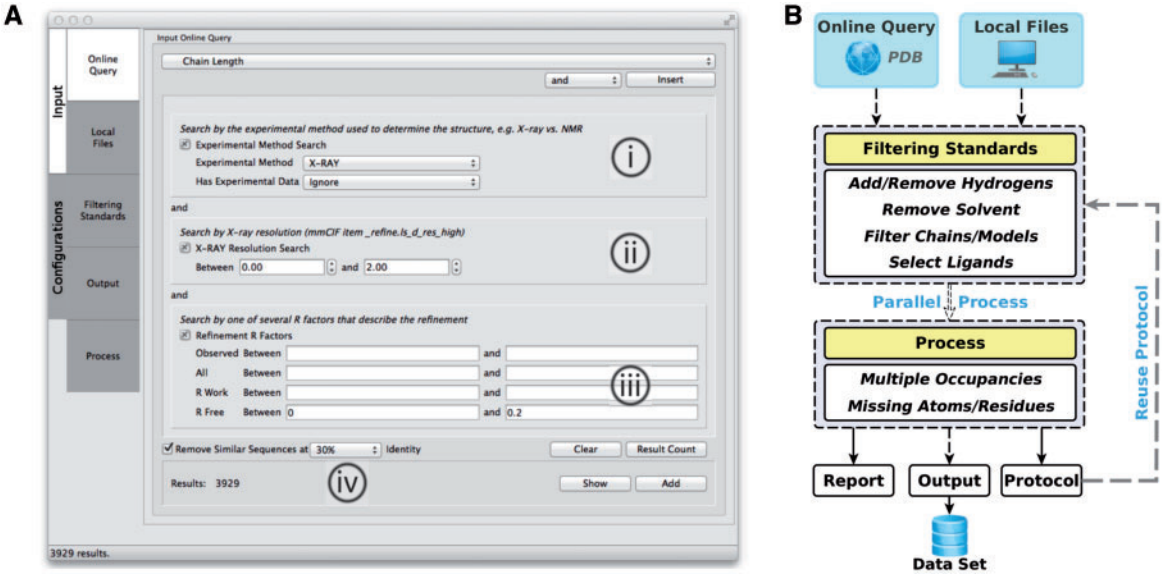

Fig. 1. PDBest GUI and workflow. (**A**) depicts the platform's main GUI features and shows the resulting screen from a PISCES-like (Wang *et al.*, 2005) online query using PDBest. The filters considered in this example were: (i) proteins solved via X-ray crystalography; (ii) X-ray resolution $\leq$ 2 Å; (iii) R-Free $\leq$ 0.2 and (iv) mutual sequence similarity lower than 30%. These filters resulted in 3929 PDB IDs, as of April 2015. The workflow in (**B**) highlights the platform's main acquisition and filtering capabilities

mutation analysis (Pires *et al.*, 2014) as well as contact network analysis (Gonçalves-Almeida *et al.*, 2012).

Funding

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); Centro de Pesquisas René Rachou (CPqRR - FIOCRUZ Minas); Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG); Financiadora de Estudos e Projetos (FINEP) and Pró-Reitoria de Pesquisa da UFMG.

Conflict of Interest: none declared.

References

Cock,P.J. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics,* **25**, 1422–1423.

da Silveira,C.H. *et al.* (2009) Protein cutoff scanning: a comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins Struct. Funct. Bioinf.*, **74**, 727–743.

Dolinsky,T.J. *et al.* (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.*, **35**, W522–W525.

Gajda,M.J. (2013) hPDB-Haskell library for processing atomic biomolecular structures in Protein Data Bank format. *BMC Res. Notes,* **6**, 483.

Gonçalves-Almeida,V.M. *et al.* (2012) HydroPaCe: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, **28**, 342–349.

Grant,B.J. *et al.* (2006) Bio3D: an R package for the comparative analysis of protein structures. *Bioinformatics*, **22**, 2695–2696.

Holland,R.C. *et al.* (2008) BioJava: an open-source framework for bioinformatics. *Bioinformatics,* **24**, 2096–2097.

Hussain,A. *et al.* (2002) PDB Goodies-a web-based GUI to manipulate the Protein Data Bank file. *Acta Crystallogr. Sect. D,* **58**, 1385–1386.

OLBoyle,N.M. *et al.* (2011) Open Babel: An open chemical toolbox. *J. Cheminform.*, **3**, 33.

Pires,D.E.V. *et al.* (2013) aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics,* **29**, 855–861.

Pires,D.E.V. *et al.* (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics,* **30**, 335–342.

Wang,G. and Dunbrack,R.L. (2005) PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.*, **33**, W94–W98.

Wiederstein,M. and Sippl,M.J. (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.*, **35**, W407–W410.