Bioinformatics Advance Access published October 13, 2016

Structural bioinformatics

PyMod 2.0: improvements in protein sequence-structure analysis and homology modeling within PyMOL

Giacomo Janson[1*], Chengxin Zhang[2], Maria Giulia Prado[1] and Alessandro Paiardini[3*]

[1]Department of Biochemical Sciences "A. Rossi Fanelli", Sapienza Università di Roma, P.le A. Moro, 5, 00185 Rome, [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA and [3]Department of Biology and Biotechnology "Charles Darwin", Sapienza Università di Roma, P.le A. Moro, 5, 00185 Rome.

*To whom correspondence should be addressed.

Associate Editor: Prof. Alfonso Valencia

Abstract

Motivation: The recently released PyMod GUI integrates many of the individual steps required for protein sequence-structure analysis and homology modeling within the interactive visualization capabilities of PyMOL. Here we describe the improvements introduced into the version 2.0 of PyMod.

Results: The original code of PyMod has been completely rewritten and improved in version 2.0 to extend PyMOL with packages such as Clustal Omega, PSIPRED and CAMPO. Integration with the popular web services ESPript and WebLogo is also provided. Finally, a number of new MODELLER functionalities have also been implemented, including SALIGN, modeling of quaternary structures, DOPE scores, disulfide bond modeling, and choice of heteroatoms to be included in the final model.

Availability: PyMod 2.0 installer packages for Windows, Linux and Mac OS X, and user guides are available at http://schubert.bio.uniroma1.it/pymod/index.html. The open source code of the project is hosted at https://github.com/pymodproject/pymod.

Contact: alessandro.paiardini@uniroma1.it; giacomo.janson@uniroma1.it

1 Introduction

Protein sequence-structure analysis (PSSA) is fundamental in a wide range of biomedical research fields, especially in protein structure prediction and modeling. Typical tasks in PSSA include similarity searches, alignments of sequences and structures, evolutionary and structural comparison, and homology modeling. While many tools are already available for the individual tasks required in PSSA (e.g., Jalview (Waterhouse *et al*., 2009) and BAM (Shapovalov *et al*., 2014)), they often do not provide integrated environments in which to seamlessly perform multiple PSSA tasks. The integration of several modeling tools into UCSF Chimera (Pettersen *et al*., 2004) was recently described (Yang *et al*., 2012). However, a similar PSSA environment for the popular molecular graphics system PyMOL (Schrödinger, LLC) has been missing until recently, when we developed PyMod 1.0 (Bramucci *et al*., 2012), a simple, yet powerful tool for sequence and structure analysis and prediction within PyMOL. While developing PyMod 1.0, we made every effort to give potential users: 1) an easy-to-learn, easy-to-use tool to fully exploit popular algorithms in PSSA, without the need to familiarize with novel graphics interfaces, or to cope with input and output file format manipulation problems; 2) full customization and control over many parameters of the individual algorithms, in order to make human knowledge-based intervention feasible during every single step of the analysis.

With these imperatives still in mind, we developed PyMod 2.0. Compared to its previous release, PyMod 2.0 has been completely rewritten and extended with a rich set of functionalities that substantially improve it, particularly in its ability to build homology models through the popular MODELLER package (Sali and Blundell, 1993). PyMod 2.0 and

PyMOL interoperate with each other and with the full complement of PyMOL plugins already available.

2 Overview of PyMod 2.0

PyMod 2.0 consists of a core sequence analysis, editing and clustering GUI extending PyMOL with an integrated toolkit of packages and web services: PSI-BLAST (Camacho et al., 2009), MUSCLE (Edgar, 2004), ClustalW (Thompson et al., 1994), Clustal Omega (Sievers et al., 2011), WebLogo 3 (Crooks et al., 2004), ESPript 3.0 (Robert and Gouet, 2014), CAMPO (Paiardini et al., 2005), PSIPRED (Jones, 1999) and MODELLER modules for models building (Sali and Blundell, 1993) and sequence-structure alignment (SALIGN; Madhusudhan et al., 2009). Users can take advantage of the implemented tools in order to perform complex tasks, such as predicting the structure of a protein target or, alternatively, make use of the individual tools for simpler PSSA tasks within PyMOL.

Template Search. This can be done: 1) by searching the online PDB database or 2) by using local PyMod databases, which can be easily updated or modified to create user-defined databases (see the PyMod manual for details). Any loaded sequence or structure can serve as a query to perform a template search, based on sequence-sequence (BLAST) or profile-sequence (PSI-BLAST) alignment methods (Altschul et al., 1997). In the latter case, users can define the number of PSI-BLAST iterations. This search generates a list of structural templates, with those most similar to the target sequence sorted at the top of the list. Users can browse the list and identify suitable templates by comparing (PSI)-Blast E-values and/or sequence identity with the target sequence. The three-dimensional structures of selected templates are retrieved and clustered with the query sequence. Users are also given the possibility to load any PDB file into PyMod and use it as template.

Sequence and Structure alignment. When two or more sequences or structures have been loaded, they can be aligned with one of the implemented sequence or structure alignment algorithms. PyMod 2.0 presents users with a rich set of algorithms for sequences and profiles alignments. Mixed structure-sequence alignments can also be built. In this scenario, structural alignments can be optionally used as guides to "merge" two or more alignments, in a way similar to that implemented in 3DCoffe (O'Sullivan et al., 2004). Additionally, PyMod 2.0 allows to easily edit alignments through an intuitive use of the mouse and to directly apply user knowledge to correct any potentially misaligned residue. Users can now identify evolutionarily conserved features in a multiple alignment with the implemented CAMPO algorithm, build distance trees of multiple sequence alignments or build sequence logos and render their alignments with the WebLogo 3 and ESPript 3.0. PyMod 2.0 has also an interface to PSIPRED, a highly accurate secondary structure predictor.

Homology Modelling. One of the main features of PyMod 2.0 is to provide an advanced GUI for many functionalities of the MODELLER package. PyMod 2.0 allows users to easily build models based on multiple structural templates, include/exclude heteroatoms and water molecules, define disulfide bridges and optimize the final model. Most importantly, homology models of entire protein complexes can be easily built, when suitable templates complexes are available. PyMod 2.0 can be used to build homology models of any kind of homo- or hetero-oligomer. To our knowledge, PyMod 2.0 is the only non-commercial tool dealing with such structural complexity in homology model building. The powerful UCSF Chimera package, for example, allows users to build models including heteroatoms, but not models of protein complexes. On the other hand, the BAM package, which does not use MODELLER as its engine, allows building models of protein complexes, but does not offer the option to include heteroatoms. The automated SWISS-MODEL server (Biasini et al., 2014) includes in models the templates heteroatoms, but only allows building of homo-oligomeric protein assemblies.

A new "refinement" window has been added to PyMod 2.0, in which the user can control the optimization of models with a number of keyword arguments (e.g., iterations steps, short MD at a fixed temperature etc.).

Structure Quality Evalulation. In order to assess the quality of a model and to highlight regions likely to contain errors, the plugin will show an interactive graph containing the DOPE (discrete optimized protein energy; Shen and Sali, 2006) profile of the model and templates. At the same time, a similar interactive Ramachandran plot graph is provided.

3 Implementation

PyMod 2.0 is written in Python and distributed under LGPL. It is compatible with PyMOL versions >= 0.99 and is available for Windows, Mac OS X and Linux.

Acknowledgements

The work of G.J. was done in partial fulfillments with the requirements of the Ph.D. degree in Biochemical Sciences at the Sapienza University of Rome. The authors are grateful to Alessandro Moretti (Sapienza University of Rome, Italy) for help. This work is dedicated to the memory of our beloved mentor Prof. Donatella Barra.

Funding

This work has been supported by grants from Sapienza University of Rome, Italy (C26A149EC4) and in part from Associazione Italiana Ricerca sul Cancro (AIRC-IG2015 n. 16720).

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Biasini,M. *et al.* (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.*, **42**, W252-258.

Bramucci,E. *et al.* (2012) PyMod: sequence similarity searches, multiple sequence-structure alignments, and homology modeling within PyMOL. *BMC Bioinformatics*, **13 Suppl 4**, S2.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Crooks,G.E. *et al.* (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Madhusudhan,M.S. *et al.* (2009) Alignment of multiple protein structures based on sequence and structure features. *Protein Eng. Des. Sel.*, **22**, 569–574.

O'Sullivan,O. *et al.* (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.*, **340**, 385–395.

Paiardini,A. *et al.* (2005) CAMPO, SCR_FIND and CHC_FIND: a suite of web tools for computational structural biology. *Nucleic Acids Res.*, **33**, W50-55.

Pettersen,E.F. *et al.* (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem*, **25**, 1605–1612.

Robert,X. and Gouet,P. (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.*, **42**, W320-324.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Schrödinger, LLC (2015) The PyMOL Molecular Graphics System, Version 1.8.
Shapovalov,M.V. *et al.* (2014) BioAssemblyModeler (BAM): user-friendly homology modeling of protein homo- and heterooligomers. *PLoS ONE*, **9**, e98309.
Shen,M.-Y. and Sali,A. (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci.*, **15**, 2507–2524.
Sievers,F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
Yang,Z. *et al.* (2012) UCSF Chimera, MODELLER, and IMP: an integrated modeling system. *J. Struct. Biol.*, **179**, 269–278.