Bramucci et al. BMC Bioinformatics 2012, 13(Suppl 4):S2
http://www.biomedcentral.com/1471-2105/13/S4/S2

BMC Bioinformatics

RESEARCH Open Access

PyMod: sequence similarity searches, multiple sequence-structure alignments, and homology modeling within PyMOL

Emanuele Bramucci[†], Alessandro Paiardini[*†], Francesco Bossa, Stefano Pascarella

From Eighth Annual Meeting of the Italian Society of Bioinformatics (BITS)
Pisa, Italy. 20-22 June 2011

Abstract

Background: In recent years, an exponential growing number of tools for protein sequence analysis, editing and modeling tasks have been put at the disposal of the scientific community. Despite the vast majority of these tools have been released as open source software, their deep learning curves often discourages even the most experienced users.

Results: A simple and intuitive interface, PyMod, between the popular molecular graphics system PyMOL and several other tools (i.e., [PSI-]BLAST, ClustalW, MUSCLE, CEalign and MODELLER) has been developed, to show how the integration of the individual steps required for homology modeling and sequence/structure analysis within the PyMOL framework can hugely simplify these tasks. Sequence similarity searches, multiple sequence and structural alignments generation and editing, and even the possibility to merge sequence and structure alignments have been implemented in PyMod, with the aim of creating a simple, yet powerful tool for sequence and structure analysis and building of homology models.

Conclusions: PyMod represents a new tool for the analysis and the manipulation of protein sequences and structures. The ease of use, integration with many sequence retrieving and alignment tools and PyMOL, one of the most used molecular visualization system, are the key features of this tool.
Source code, installation instructions, video tutorials and a user's guide are freely available at the URL http://schubert.bio.uniroma1.it/pymod/index.html

Background

Once confined only to experts in bioinformatics, protein sequence retrieving, aligning and modeling tasks are now being routinely approached by an increasing number of researchers, who can take also advantage of the growing number of structures that are being deposited every day in public databases. Integrating protein sequence and structure information has therefore become an imperative, especially in the field of protein structure prediction from sequence, by means of homology modeling (HM) methodologies.

In recent years, a number of valuable tools related to protein sequence analysis and modeling (e.g., DeepView [1], MolIDE [2] and Chimera [3]) has been developed. While these tools are in many cases easily accessible, and have greatly simplified some of the problems that are most frequently encountered when coping with sequence/structure analysis tasks (e.g., lack of graphical user interfaces [GUIs], need to make use of many programs in an integrated way and input and output file format manipulation problems), the initial difficulties and deep learning curves often encountered when mastering the usage of new software sometimes discourages first-time, as well as more experienced users. On the other hand, public servers (e.g., Phyre [4], CPHmodels [5]), which are able to automatize some or all of the

* Correspondence: alessandro.paiardini@uniroma1.it
† Contributed equally
Dipartimento di Scienze Biochimiche "A. Rossi Fanelli", Sapienza Università di Roma, Roma 00185, Italy

main modeling tasks, often do not offer users the ability to apply knowledge-based intervention during the analysis (e.g., sequences selection, manual refinement of multiple alignments and choice of parameters during model construction).

In order to contribute to tackle these issues, a simple and intuitive interface between the open-source and widely used biomolecular visualization program PyMOL [6] and several other well-known sequence/structure analysis tools (i.e., BLAST [7], PSI-BLAST [8], MUSCLE [9] ClustalW [10], CEalign [11] and MODELLER [12]; Table 1), has been developed. The tool presented here, PyMod, aims to give researchers and students with no or a limited familiarity in this field, as well as more experienced users, the ability to exploit popular algorithms in sequence/structure analysis and protein structure prediction, and most importantly full customization and control over their parameters, while retaining as much as possible an ease of use and the familiarity of the PyMOL environment (Figure 1).

Implementation

PyMod has a rich functionality, based on its core sequence alignment, clustering and editing window. These features are described in outline in the following sub-sections.

Similarity searches

PyMod can input and output sequences and 3D-structures in the popular FASTA and PDB formats. In the latter case, 3D-coordinates are automatically split in single chains, loaded into PyMOL, and their corresponding sequences loaded into the PyMod main window (Figure 2). After a sequence has been loaded onto the PyMod main window, users can search different databases, in order to retrieve protein sequences and related structures that are homologous to the query sequence, by means of the BLAST and PSI-BLAST search tools. BLAST is relatively faster while less sensitive when compared with profile-profile alignment methods. However, it can still detect homology with significant sequence identity (i.e., identity > 40%) [8,13,14], thus providing fast and useful means in the case of high identity, template-based modeling. On the other hand, PSI-BLAST, the most used profile-sequence alignment method, is more sensitive than sequence-sequence alignment and it can recognize distant homology with lower sequence identity (i.e., identity > 20%) [8]. Both tools have been therefore implemented in PyMod. Profile-profile alignments or HMM-HMM (Hidden Markov Models) comparison algorithms [15] may be the most effective approaches and even able to create accurate alignments in extreme cases (i.e., identity < 10%) [16], but they're usually much more complex and slower than sequence-sequence or profile-sequence alignments. Most notably, at these levels of sequence identity (0-20%), fold-recognition or *ab initio* approaches may be favored over homology modeling, for which PyMod flowchart has been primarily planned. PyMod includes support for running BLAST remotely (no local database installation is required) and PSI-BLAST locally. In the latter case, users are provided with the option to install local sequence databases, while PyMod provides a graphical interface to ease their use. To facilitate template structure search for homology modeling tasks, PyMod will be distributed with a pre-installed PDB sequence database, which will be updated in future releases on a monthly base. A number of (PSI-)BLAST parameters can be controlled by the user from within an apposite PyMod window (e.g., number of PSI-BLAST iterations, E-value threshold, % identity threshold) (Figure 3). Users are provided with the ability to select the (PSI-)BLAST results to be imported in the PyMod main window, by choosing from a table reporting the name of the retrieved sequences, their E-value and sequence identity. Selected sequences, once imported in PyMod, are automatically grouped in a separate cluster, which can be collapsed or contracted by simply clicking a button beside the query sequence. When searching the PDB database the user can retrieve the 3D-coordinates that are related to a selected query and automatically load the structure into the PyMOL main window.

As such, PyMod provides a graphical interface for (PSI-)BLAST searches of large databases, both locally or remotely, which can be also used as a standalone tool inside the PyMOL framework.

Alignment of sequences and structures

Once retrieved sequences from selected databases are loaded in PyMod, they can be used to generate a

Table 1 PyMod integrated tools

BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi
PSI-BLAST	http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&PAGE=Proteins&PROGRAM=blastp&RUN_PSIBLAST=on
MUSCLE	http://www.drive5.com/muscle/
ClustalW	http://www.ebi.ac.uk/Tools/msa/clustalw2/
Cealign	http://cl.sdsc.edu/ce.html
Modeller	http://salilab.org/modeller/

This table summarizes the tools that have been integrated into PyMod providing their URLs.

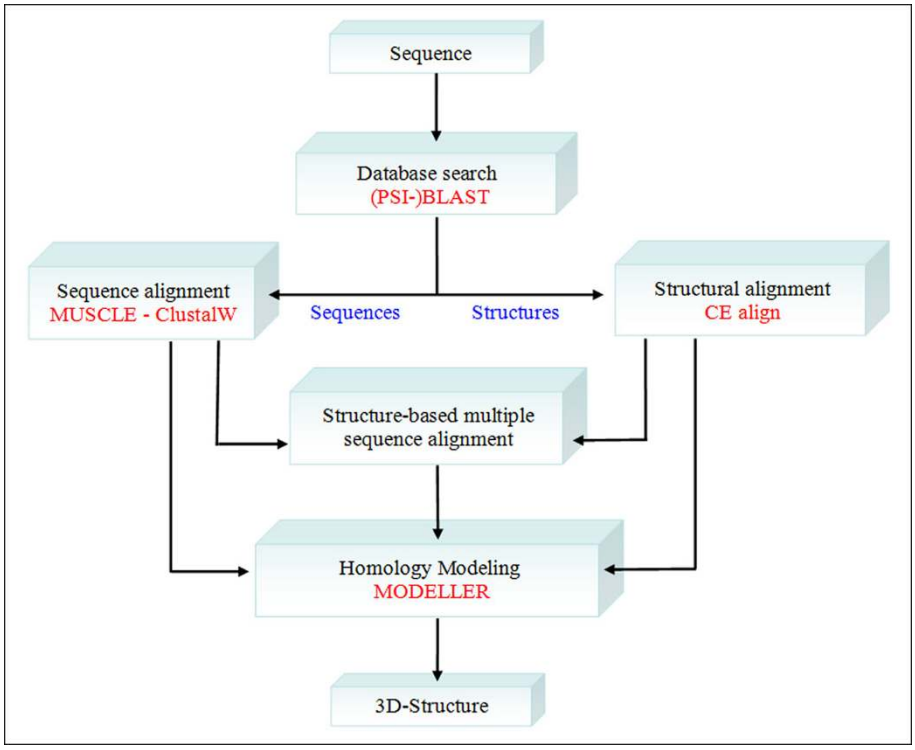

Figure 1 PyMod integrated tools flowchart. The workflow shows how the separate tools are integrated in PyMod. Each tool must be considered as *standalone* (e.g., it's possible to perform a sequence alignment task without searching in the database as a mandatory step).

multiple sequence alignment by means of MUSCLE and ClustalW programs. The choice of two multiple sequence alignment tools is twofold: on the one hand, ClustalW is famous and very popular among people with limited experience in the field; on the other hand, MUSCLE is known to outperform ClustalW in quality and in speed. Future implementation of additional tools (e.g., T-Coffe [17]) is planned. Additionally, PyMod can input and output multiple alignments in the popular FASTA and Clustal formats. Different multiple alignments can be built for each cluster of sequences that is available in PyMod. The user from within an apposite PyMod window can control a number of ClustalW parameters. When dealing with sequence alignments comprising known 3D-structures, it is always more desirable to exploit this kind of information, by performing structural superposition and deriving a structure-based alignment. In this case, users can carry out a multiple structural alignment by using the combinatorial extension algorithm, implemented in the popular program CE, a fast and robust algorithm in superposing and aligning 3D-structures [11]. The selected 3D-structures are then automatically superposed in PyMOL, and the resulting structural alignment is displayed in PyMod. If the 3D-structures to be superposed and aligned have been previously aligned to their own sequence cluster with MUSCLE or ClustalW, users can optionally keep the latter, by using the structural alignment as guide to "merge" the two alignments. In this way, structure-based alignments are used as a template for realigning the

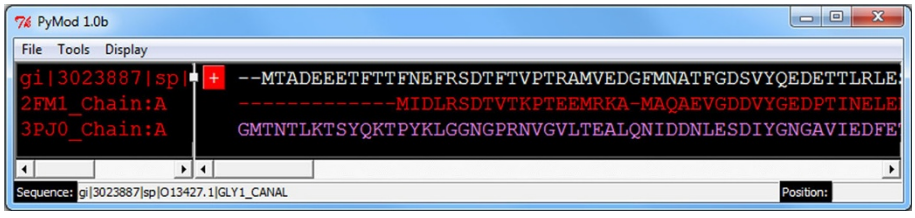

Figure 2 PyMod main window. By using the main window, the user can perform almost all PyMod functions using the top menu or just operating on the sequences and their name.

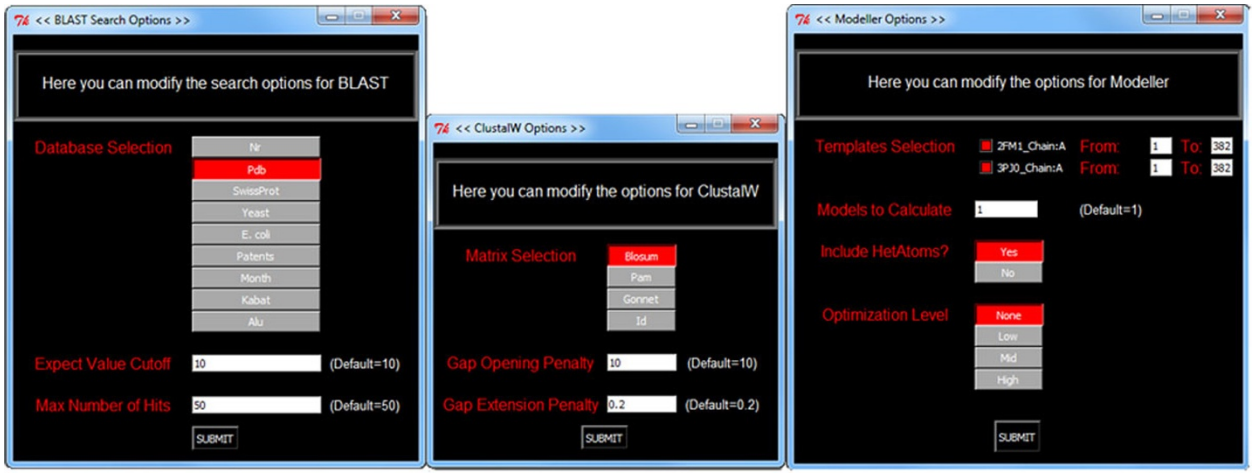

Figure 3 Preferences window. The user can change many parameters of the tools implemented in PyMod through specific Options windows.

original sequences, obtaining a structure-based multiple sequence alignment that combines sequences and structures. This procedure is similar to the one already implemented in the 3DCoffe tool [18]. The option to generate mixed structure-sequence alignment is particularly useful when two or more evolutionarily distant structural templates and their close orthologous sequences have to be aligned. In this scenario, the structural alignment of templates (which, being based on structure superposition, would outperform any sequence-based method) will provide the starting point for the subsequent merging of orthologous sequences, which have been previously aligned with MUSCLE or ClustalW. Most importantly, manually editing multiple sequence alignments in PyMod will allow the user to apply her/his knowledge to correct any misaligned residue. This option in PyMod simply requires the user to click with the mouse at the desired sequence position and then to drag residues to the right/left to add/remove gaps. The ability to edit sequences is another feature implemented in PyMod. Indeed, during modeling tasks, it is often necessary to mutate and/or trim existing sequences at their ends. This option, for example, helps to prevent long overhanging fragments after a MODELLER run. Excising part of the sequence in the middle is also possible. Finally, a number of coloring options for sequences are available via the PyMod menu, including a secondary structure scheme for sequences related to 3D-structures (Figure 4).

Homology modeling

Starting from a previously obtained alignment, it is possible to build a homology model of a selected sequence through the PyMod interface to the popular MODELLER program. The satisfaction of spatial restraints algorithm, as implemented in MODELLER, undoubtedly represent one of the most popular homology modelling approaches, and has become the model-building program of choice for several homology modelling servers because of its relative speed and reliability. Several of the strongest performing prediction servers in the CASP8 experiment, such as HHpred [19], incorporate MODELLER in their methodology. When compared against other homology modeling programs MODELLER is considered one of the better performing structure predictors [20]. Up till now, there have been a few attempts earlier to simplify the use of MODELLER by providing a GUI framework (EasyModeller [21], SWIFT MODELLER [22]).

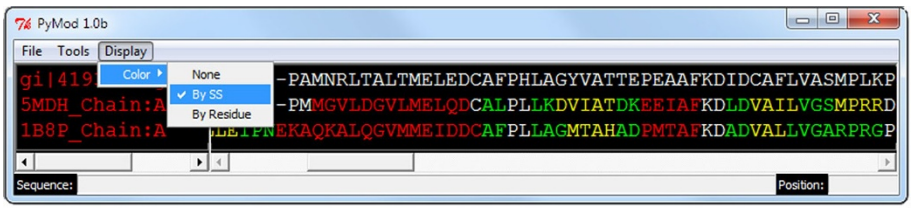

Figure 4 Secondary structure colouring method. The user can color the protein sequences by their secondary structure or by the physicochemical properties of their amino acids. The secondary structure assignment is based on the proSS algorithm http://roselab.jhu.edu/utils/pross.html.

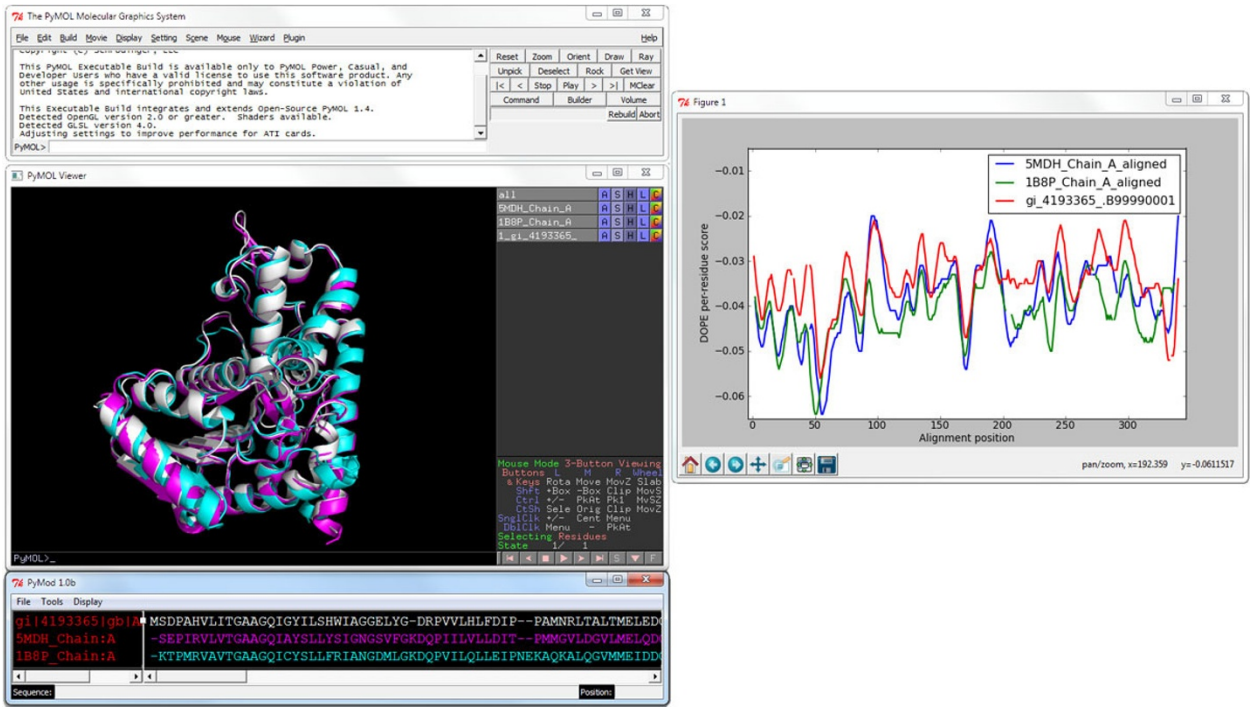

Figure 5 Homology model and quality assessment. Once homology models have been built, they are automatically loaded onto PyMOL and the user can assess the models quality through a graph constructed using the DOPE scoring function.

Merging MODELLER with the most popular tools for sequence retrieving and sequence/structure alignments ([PSI-]BLAST, ClustalW, MUSCLE, CE), within the PyMOL framework gives an unprecedented level of ease and control over all of the tasks required to construct homology models with MODELLER. A number of MODELLER parameters can be controlled by the user from within an apposite PyMod window. These include the choice of structural template(s), the model refinement level, the number of models to build, and the specification of the region to be modeled. By merging the versatility of MODELLER and the user-friendly PyMod/PyMOL environment, it is also possible to easily include heteroatoms (e.g., inhibitors, docked substrates, cofactors) in the final models, a pivotal feature that is often absent in many state of the art tools for homology modeling. Finally, the pipeline includes a validation tool (DOPE, or Discrete Optimized Protein Energy [23]), to highlight regions where the alignment/model is accurate and where it is likely to be incorrect (Figure 5). Once homology models have been built, they are automatically loaded onto PyMOL for visual inspection and further analysis.

Conclusions

PyMod represents a new tool for the analysis and the manipulation of protein sequences and structures. The ease of use, integration with many sequence retrieving and alignment tools and PyMOL, one of the most used molecular visualization system, are the key features of this tool. We plan to release future updates of PyMod, including additional tools for secondary structure prediction, sequence retrieving and alignment, as well as other tools suggested by the users' community. Finally, a tighter integration between PyMOL, MODELLER and PyMod will constitute a main issue of future project development plans.

Availability and requirements

Project name: PyMod

Project home page: http://schubert.bio.uniroma1.it/pymod

Operating system(s): Windows (XP, Vista, Seven). Linux (Ubuntu) and Mac OS (10.6) will be supported in the next release.

Programming language: Python

License: Lesser General Public License (LGPL)

Other requirements: PyMOL version 1.1.1 or newer, BioPython version 1.50 or newer, Standalone BLAST 2.2.25+ or newer, Muscle, ClustalW and MODELLER.

Acknowledgements

This work was partially supported by the funds of the Italian "Ministero dell'Istruzione, dell'Università e della Ricerca" and by the "Consorzio Interuniversitario per le Applicazioni di Supercalcolo per Università e Ricerca." (CASPUR, Roma, Italy) [std11-459]. This work will be submitted by EB in

partial fulfillment of the requirements of the degree of "Dottorato di Ricerca in Biochimica" at Sapienza, Università di Roma.
This article has been published as part of *BMC Bioinformatics* Volume 13 Supplement 4, 2012: Italian Society of Bioinformatics (BITS): Annual Meeting 2011. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/13/S4.

Authors' contributions

EB wrote the software and helped to draft the manuscript. AP conceived the study, helped to write the program and drafted the manuscript. FB revised the manuscript critically for important intellectual content. SP participated in the study design and coordination and revised the manuscript. All authors read and approved the final manuscript. EB and AP contributed equally to this work.

Competing interests

The authors declare that they have no competing interests.

Published: 28 March 2012

References

1. Arnold K, Bordoli L, Kopp J, Schwede T: **The SWISS-MODEL Work-space: a web-based environment for protein structure homology modelling.** *Bioinformatics* 2006, **22**:195-201.
2. Canutescu AA, Dunbrack RL Jr: **MolIDE: a homology modeling framework you can click with.** *Bioinformatics* 2005, **21**:2914-2916.
3. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: **UCSF Chimera–a visualization system for exploratory re-search and analysis.** *J Comput Chem* 2004, **25**:1605-1612.
4. Kelley LA, Sternberg MJE: **Protein structure prediction on the web: a case study using the Phyre server.** *Nature Protocols* 2009, **4**:363-371.
5. Nielsen M, Lundegaard C, Lund O, Petersen TN: **CPHmodels-3.0 - Remote homology modeling using structure guided sequence profiles.** *Nucleic Acids Research* 2010, **38**:W576-W581.
6. DeLano WL: **The PyMOL Molecular Graphics System.** San Carlos, CA: DeLano Scientific; 2002.
7. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
8. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
9. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**:1792-1797.
10. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**:4673-4680.
11. Shindyalov IN, Bourne PE: **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.** *Protein Eng* 1998, **9**:739-747.
12. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A: **Comparative protein structure modeling using MODELLER.** *Curr Protoc Bioinformatics* 2006, **Chapter 5**:Unit 5.6.
13. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *Journal of Molecular Biology* 1970, **48**:443-453.
14. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *Journal of Molecular Biology* 1981, **147**:195-197.
15. Remmert M, Linke D, Lupas AN, Soding J: **HHomp–prediction and classification of outer membrane proteins.** *Nucleic Acids Res* 2009, **37**: W446-W451.
16. Yona G, Levitt M: **Within the Twilight Zone: a sensitive profile-profile comparison tool based on information theory.** *J Mol Biol* 2002, **315**:1257-1275.
17. Notredame C, Higgins D, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *Journal of Molecular Biology* 2000, **302**:205-217.
18. O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C: **3DCoffee: combining protein sequences and structures within multiple sequence alignments.** *J Mol Biol* 2004, **340**:385-395.
19. Söding J, Biegert A, Lupas AN: **The HHpred interactive server for protein homology detection and structure prediction.** *Nucleic Acids Res* 2005, **33(Web Server issue)**:W244-W248.
20. Wallner B, Elofsson A: **All are not equal: a benchmark of different homology modeling programs.** *Protein Sci* 2005, **14**:1315-27.
21. Kuntal BK, Aparoy P, Reddanna P: **EasyModeller: a graphical interface to MODELLER.** *BMC Res Notes* 2010, **3**:226-330.
22. Mathur A, Shankaracharya , Vidyarthi AS: **SWIFT MODELLER: A JAVA based GUI for molecular modeling.** *J Mol Model* 2011, **17**:2601-2607.
23. Shen M-y, Sali A: **Statistical potential for assessment and prediction of protein structures.** *Protein Science* 2006, **15**:2507-2524.

doi:10.1186/1471-2105-13-S4-S2
Cite this article as: Bramucci *et al.*: PyMod: sequence similarity searches, multiple sequence-structure alignments, and homology modeling within PyMOL. *BMC Bioinformatics* 2012 **13**(Suppl 4):S2.