W550–W554 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* *Published online 4 June 2010*
doi:10.1093/nar/gkq475

PUDGE: a flexible, interactive server for protein structure prediction

Raquel Norel*, Donald Petrey and Barry Honig*

Department of Biochemistry and Molecular Biophysics, Howard Hughes Medical Institute, Center for Computational Biology and Bioinformatics, Columbia University, 1130 St. Nicholas Avenue, New York, NY 10032, USA

Received January 31, 2010; Revised May 6, 2010; Accepted May 13, 2010

ABSTRACT

The construction of a homology model for a protein can involve a number of decisions requiring the integration of different sources of information and the application of different modeling tools depending on the particular problem. Functional information can be especially important in guiding the modeling process, but such information is not generally integrated into modeling pipelines. Pudge is a flexible, interactive protein structure prediction server, which is designed with these issues in mind. By dividing the modeling into five stages (template selection, alignment, model building, model refinement and model evaluation) and providing various tools to visualize, analyze and compare the results at each stage, we enable a flexible modeling strategy that can be tailored to the needs of a given problem. Pudge is freely available at http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:PUDGE.

INTRODUCTION

Computational modeling of protein structures has become an effective means of generating new biological information. Ideally, a structure prediction server would provide a simple, sequence in/structure out approach that generates a single 'best' model for any given protein sequence. However, there is a wide variety of tools available that carry out different stages of the process, and in general no one set of tools will be applicable to every problem. Moreover, the definition of what constitutes a best model can be ambiguous and will depend on factors such as sequence/structural features of the available templates and their degrees of homology to the target. Modeling can be much more effective if function is taken into account, but servers have not generally been designed explicitly to put the modeling process in a biological context.

Pudge is a protein structure prediction server designed to address these issues. The overall modeling strategy implemented in Pudge is illustrated in Figure 1 (the individual features of Pudge are described throughout the text as well as in the figure caption). A number of sequence-based calculations are carried out when the sequence is initially submitted. These include secondary structure prediction (1) the identification of low-complexity (2) and disordered (3) regions, as well as the optional splitting of the sequence into domains (2). The modeling process is then divided into a pipeline consisting of five stages: template selection (TS), alignment (AL), model building (MB), model refinement (MR) and model evaluation (ME). Menu options allow a user to apply different methods at each stage. The methods that we have implemented have been chosen to provide a wide range of flexibility in what sequence and structural relationships to make use of.

Specifically, simple BLAST runs or more sophisticated profile–profile approaches can be applied to identify modeling templates. At the AL stage, a single alignment to different templates or alternate alignments to a particular template can be generated. One or up to five alternate models per template can be constructed at the MB stage. Side chains and loops can be further sampled using methods that apply fast empirical force fields or all-atom potentials at the MR stage. Finally, a set of statistical potentials can be applied to evaluate and rank the set of models generated at the ME stage. A complete list of methods with appropriate references is provided in Table 1.

An important feature of the Pudge server is its interactivity. Initially, it can often be unclear, which templates, alignments or modeling methods are appropriate to a particular problem. To address this, Pudge provides, in addition to the modeling methods themselves, a set of methods to analyze and compare the results produced at each stage. Based on the results of the analysis,

*To whom correspondence should be addressed. Fax: +1 212 851 4650; Email: bh6@columbia.edu
Correspondence may also be addressed to Raquel Norel. Tel: +1 212 815 4655; Email: rn98@columbia.edu

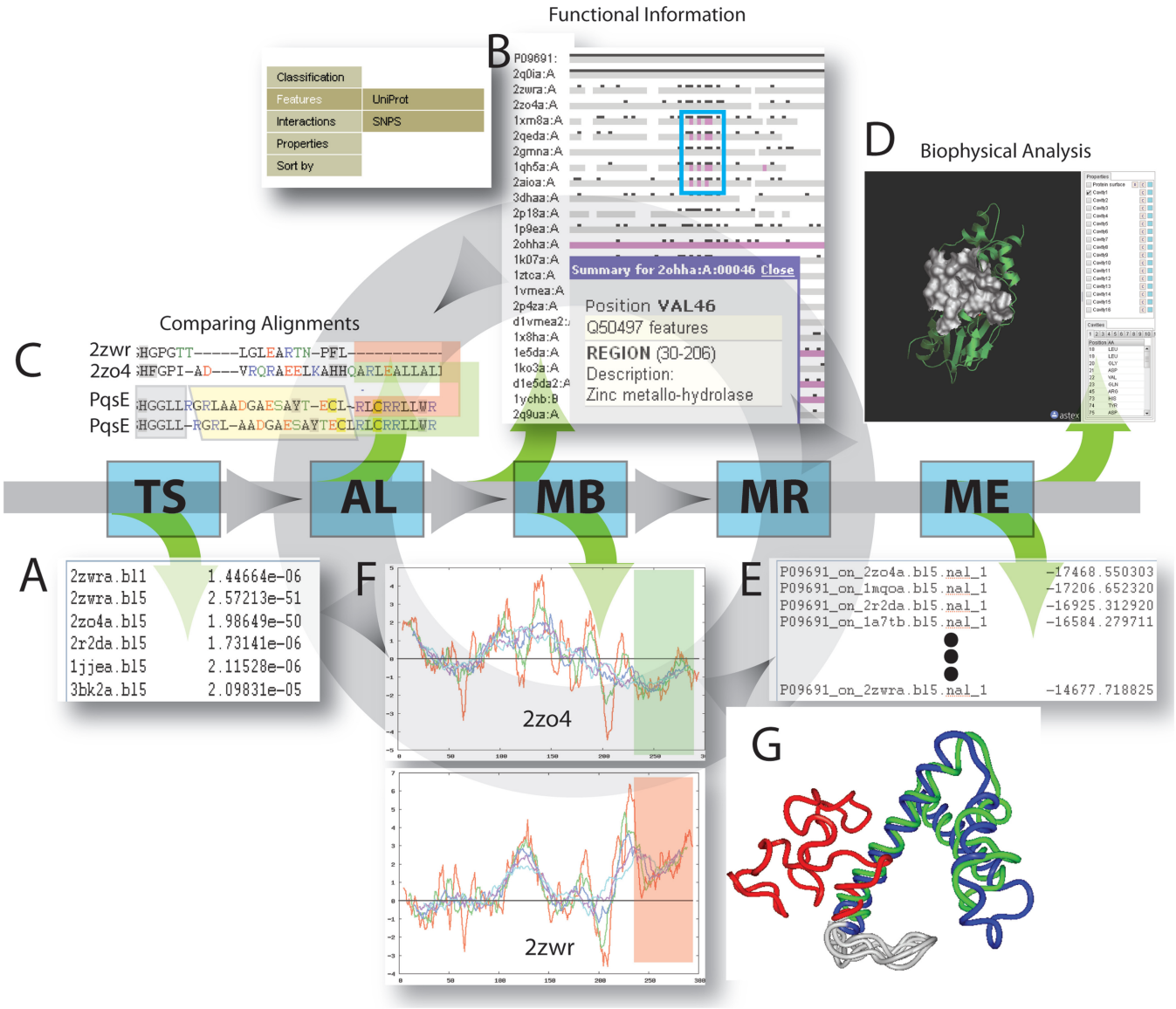

Figure 1. The Pudge protein structure prediction pipeline. Starting with a sequence, a model is constructed in five stages: TS, AL, MB, MR and ME. Menu options allow a user to select the methods applied at each stage but models can also be generated in an entirely automated fashion. Results from each stage can be examined and edited on the web page (**A**, **E**). A variety of tools can be applied to analyze the edited results. Those illustrated here are taken from a case study modeling the protein PqsE described in the text. (**B**) The 'annotation map'. Functional properties of all the templates can be examined, queried and filtered using the annotation map. A wide range of properties can be displayed (see text). For example, the blue rectangle highlights residues in metal-binding active sites (magenta squares), which are conserved in PqsE (black bars). (**C**). Simultaneous comparison of the alignments of PqsE to two templates (PDB codes 2zwr and 2zo4). The figure indicates that the models built on these templates would be equivalent in the gray region and would have an alignment shift in the yellow region. Models built using 2zwr in the red region would be built *ab initio*, whereas that region is structured and well-aligned for the template 2zo4 (green, see also Figure 1G). (**D**) Any individual model can be submitted to our MarkUs protein function annotation server for additional bioinformatic and biophysical analysis. The figure shows a molecular viewer within MarkUs displaying predicted functional cavities. (**F**) A ProSa2003 evaluation of models of PqsE built on templates 2zo4 and 2zwr. The C-terminal of the model built on 2zwr has lower quality (red rectangle) than the model built on 2zo4 (green rectangle). (**G**) Worm representations of the C-termini of the native structure of PqsE, and the models built on 2zwr and 2zo4. The models were largely equivalent (e.g. gray regions) but the C-terminal of 2zo4 (green) was a close structural match to the PqsE native structure (blue), whereas the same region had to be built *ab initio* in the model built on 2zwr (red) even though PqsE was more similar to this template overall.

individual templates, alignments and models can be resubmitted to the pipeline for additional processing. An important feature of the server is the ability to conveniently incorporate functional information into the analysis of what should be done at each stage. This is done via an interface to our protein function annotation server, MarkUs (http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:Mark-Us). The MarkUs (4) interface provides unique functionality that allows a user to examine, query and compare, functional information for a set of templates over a range of similarities, facilitating the identification of individual sequence and structural features that may be important both for guiding the modeling and accurately determining the biological role of the query protein.

Table 1. Modeling and analysis methods used in the PUDGE pipeline

TS	AL	MB	MR	ME
BLAST (5,6) HMAP (7,8)	BLAST (5,6) Smith–Waterman (18) HMAP(7,8)	NEST (7,9) MODELLER (19)	SCAP (9,20) LOOPY (21) PLOP (20) SCWRL(22)	SVM (Zhu and Honig, manuscript in preparation) DFIRE (12) Verify3D (10,11) ProSa2003 (16)

All methods used have been independently validated and appropriate references are cited. Additional details on how the methods are specifically used within Pudge are provided as 'tool tips' on the Pudge web site itself.

A CASE STUDY: PqsE

We describe the use of the server by modeling the protein PqsE, a member of an operon involved in quorum sensing in *Pseudomonas aeruginosa*. The example uses only a sampling of the tools available in the server, but illustrates some common difficulties that arise in any modeling exercise, and highlights the need to examine and compare results from different stages of the process and to integrate functional information in order to generate a more accurate and useful model. Since in general it is not known initially what modeling templates are available or what their ranges of similarity are to the query, Pudge provides a simplified two-step interface to generate a number of models using a default set of methods. These methods are intended to provide results from a variety of tools that exploit a wide spectrum of sequence relationships, from close to remote, and generally provide a set of results that is sufficient to make initial decisions that guide the modeling.

Currently our default methods include a 1-iteration PSI-Blast (5,6) run to identify close homologs, a 5-iteration PSI-Blast run and a search using our in-house profile–profile comparison method, Hybird Multidimensional Alignment Profile (HMAP) (7,8) to identify more remote templates. Templates are selected from a 90% non-redundant database of single protein chains, which is updated weekly. Sequence-to-structure alignments to the templates selected by all three methods are also generated using HMAP. Models are built using our program NEST (7,9), and are evaluated using the Verify3D (10,11) and DFIRE (12) statistical potentials. The methods that were applied are displayed on a web page as sets of check boxes. The models, templates and alignments generated by each method can be examined by checking these boxes, which generates lists specifying an identifier for the target sequence, PDB codes of the templates and shorthand labels identifying the methods used (e.g. Figure 1A and E). These lists can be edited, removing unessential or non-promising solutions. Using menu options on the web page, the edited list of templates, alignments or models can be resubmitted to the pipeline for additional analysis or modeling.

Analyzing function

An important method implemented as a default is an analysis of the 'coverage' of the query sequence by the templates, which is carried out at the AL stage. This analysis allows examination of functional information associated with the templates selected by the different TS methods. The information is visualized using a feature of our MarkUs protein function annotation server called the annotation map, which represents the query and template sequences in a web-based 'BLAST-like' format (Figure 1B). This visualization includes many features that allow examination, querying and filtering of information from a wide range of functional databases and is a useful place to start establishing a biological context for the modeling.

For example, simple browsing of overall annotations from different sources [Gene Ontology (GO), Enzyme Classification (EC) and UniProt] can be carried out and, in our example, examination of the overall annotations of the closest sequence neighbors of PqsE identifies it as metallo-hydrolase. Transferring annotations based on the overall sequence similarity can always be problematic, however, and it is generally necessary to confirm an annotation by ensuring that specific sequence or structural features responsible for activity are conserved. Menu options allow a user to display residue-specific functional information in the annotation map [e.g. UniProt sequence features, Single Nucleotide Polymorphism (SNPs) and protein–ligand interactions]. In the case of PqsE, mapping UniProt features both confirms the metallo-hydrolase annotation based on conservation of metal-chelating residues and identifies the active site, which will be useful in interpreting the modeling results. Other types of manipulations can also be carried out. For example, if a more likely specific function is known (e.g. if a specific substrate is suspected) the set of templates can be limited to those with similar functional features based on GO annotation to ensure that the templates are functionally related. Similar options can be used to identify those templates that bind specific ligands or types of ligands based on the ChEBI (13–15) annotations available in the PDB (a tutorial and general description for the MarkUs features is available at http://luna.bioc.columbia.edu/honiglab/nesg/documentation/tutorial.html).

Analyzing models and alignments

While it is generally reasonable to expect that an optimal model can be generated based on the closest sequence homolog (perhaps identified using a single iteration of PSI-Blast), there are often exceptions to this rule and a careful comparison of results generated by the default methods is often useful. In the case of PqsE, differences between the sets of templates selected using different

criteria at different stages of the pipeline can give important clues about what templates may be most appropriate to use. A single iteration of BLAST identifies only a single close homolog (2zwr; Figure 1A). Additional iterations also identify 2zwr as the 'best' template but also more remotely related proteins, including 2zo4 with an *E*-value essentially equivalent to that of 2zwr (at least for the purposes of modeling; Figure 1A). Significantly, an examination of the results at the ME stage reverses this trend, with the model based on 2zo4 now ranked as the most favorable in terms of the statistical potential DFIRE (12) and the model based on 2zwr ranked poorly (22nd among the models evaluated).

When different measures give conflicting information, it is useful to identify the source of the difference. In the commonly occurring situation where one is dealing with ambiguous templates, Pudge provides several tools that allow a user to identify problematic regions of the models. Variations such as those encountered here for PqsE typically occur for one of two reasons: (i) the alignments to the two templates do not produce equivalent models; or (ii) there are structural differences between the two templates. To identify alignment differences, a user edits the results at the AL stage and chooses the 'Compare alignments' tool from the menu option provided. The output is an alignment (Figure 1C) that includes the sequences of all the templates (aligned structurally) as well as multiple copies of the query sequence (details of how it is constructed are provided in the tutorial on the web site). The alignment helps to identify regions where the query sequences do not align perfectly. For the case of PqsE, one of these regions occurs at the C-terminal, where the output of 'Compare alignments' indicates that 2zwr is not long enough to model PqsE completely even though it is a better match in terms of overall sequence similarity (Figure 1C).

On the other hand, the template 2zo4 is sufficiently long, but this is no guarantee that it is an accurate structural match for PqsE. This issue can be further studied by comparing the modeled structures of PqsE themselves. This is carried out by again editing the results list as in Figure 1A (but in this case for the MB step), and selecting the 'Resubmit' menu option, which presents a user with a set of methods to analyze, refine and evaluate the selected models. In our example, we resubmitted the models constructed using the templates 2zwr and 2zo4 and applied the ProSa2003 (16) residue-by-residue ME method. The output is shown in Figure 1F and clearly indicates a problem with the model based on 2zwr (positive ProSa scores at the C-terminal). On the other hand, the ProSa scores of the model based on 2zo4 are highly favorable in this region and in fact the model based on 2zo4 contains the correct conformation of the C-terminal (Figure 1G), as judged by a comparison to the native structure (there are other areas of variation as well, which are ignored here for simplicity). This issue is especially important when the models are considered in biological context, since the C-terminal helices form tunnels leading to the active site (which was identified using the annotation map), a feature which has been suggested to indicate that the protein acts on extended substrates (17).

Once issues such as those described above have been addressed and a satisfactory model has been constructed, Pudge provides a mechanism to submit individual models to our protein function annotation server, MarkUs. While the coverage analysis described above provides access to a great deal of functional information, it will necessarily be limited to those proteins that have some sequence similarity to the query. MarkUs carries out a more detailed bioinformatic and biophysical analysis of the model, which includes the identification of putative functional cavities, conservation in the target sequence and electrostatics, all of which can be visualized in a molecular viewer (Figure 1D). A more comprehensive examination of functional information associated with structurally similar proteins is also carried out, which can be examined in a separate annotation map. Coarse modeling of protein–protein, protein–DNA and protein–ligand interactions is also possible.

CONCLUSIONS AND FUTURE DIRECTIONS

Ideally, a set of modeling tools would be available that integrates different sources of functional and structural information and automatically generates an optimal model. In practice, however, there is no general computational formula that invariably recognizes and accurately deals with issues such as those described here, and a researcher interested in a detailed and reliable understanding of what a model may indicate about function (and vice versa) needs at least to be aware of what the issues are. Pudge is designed to be a convenient interface for this type of analysis.

The initial analysis of templates, models and alignments has been our primary focus here, but an important question is what to do once specific modeling issues have been identified. The accuracy needed for a model depends on how it is going to be used and the information garnered from the results described above (i.e. PqsE is a likely metallo-hydrolase with a likely C-terminal capping region that governs access to the active site) may be sufficient to suggest possible substrates or further guide experiments. If a more accurate model is needed, there are a number of tools available, but which strategy to use is not always clear and will depend on the expertise of the user. While Pudge contains many options to further sample alignments and refine and evaluate models as a whole (Table 1), an optimal model in the case of PqsE would probably need to be based on a combination of alignments and structures of the two templates described in the above analysis. An important future development goal of the server is to provide tools that allow for finer selection over which templates and alignments (or combinations thereof) to use at each stage of the process.

FUNDING

This work was supported by the National Institute of Health (grant numbers GM030518, GM074958, CA121852). Funding for open access charge: Howard Hughes Medical Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
2. Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.
3. Dosztanyi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.
4. Petrey,D., Fischer,M. and Honig,B. (2009) Structural relationships among proteins with different global topologies and their implications for function annotation strategies. *Proc. Natl Acad. Sci. USA*, **106**, 17377–17382.
5. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
6. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
7. Petrey,D., Xiang,Z., Tang,C.L., Xie,L., Gimpelev,M., Mitros,T., Soto,C.S., Goldsmith-Fischman,S., Kernytsky,A., Schlessinger,A. *et al*. (2003) Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins*, **53(Suppl. 6)**, 430–435.
8. Tang,C.L., Xie,L., Koh,I.Y., Posy,S., Alexov,E. and Honig,B. (2003) On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. *J. Mol. Biol.*, **334**, 1043–1062.
9. Xiang,Z. and Honig,B. (2001) Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.*, **311**, 421–430.
10. Bowie,J.U., Luthy,R. and Eisenberg,D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.
11. Luthy,R., Bowie,J.U. and Eisenberg,D. (1992) Assessment of protein models with 3-dimensional profiles. *Nature*, **356**, 83–85.
12. Zhou,H. and Zhou,Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, **11**, 2714–2726.
13. de Matos,P., Alcantara,R., Dekker,A., Ennis,M., Hastings,J., Haug,K., Spiteri,I., Turner,S. and Steinbeck,C. Chemical entities of biological interest: an update. *Nucl. Acids Res.*, **38**, D249–254.
14. Degtyarenko,K., de Matos,P., Ennis,M., Hastings,J., Zbinden,M., McNaught,A., Alcantara,R., Darsow,M., Guedj,M. and Ashburner,M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucl. Acids Res.*, **36**, D344–350.
15. Degtyarenko,K., Hastings,J., de Matos,P. and Ennis,M. (2009) ChEBI: an open bioinformatics and cheminformatics resource. *Curr. Protoc. Bioinformatics*, **Chapter 14**, Unit 14.19.
16. Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355–362.
17. Yu,S., Jensen,V., Seeliger,J., Feldmann,I., Weber,S., Schleicher,E., Häussler,S. and Blankenfeldt,W. (2009) Structure elucidation and preliminary assessment of hydrolase activity of PqsE, the Pseudomonas quinolone signal (PQS) response protein. *Biochemistry*, **48**, 10298–10307.
18. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
19. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
20. Jacobson,M.P., Friesner,R.A., Xiang,Z. and Honig,B. (2002) On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.*, **320**, 597–608.
21. Xiang,Z., Soto,C.S. and Honig,B. (2002) Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc. Natl Acad. Sci. USA*, **99**, 7432–7437.
22. Canutescu,A.A., Shelenkov,A.A. and Dunbrack,R.L. Jr (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.*, **12**, 2001–2014.