Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W369–W374***
doi:10.1093/nar/gkm319

Pcons.net: protein structure prediction meta server

Björn Wallner[1,*], Per Larsson[2] and Arne Elofsson[2]

[1]Department of Biochemistry, University of Washington, Box 357350, Seattle, WA 98195, USA and
[2]Center for Biomembrane Research, Stockholm University, SE-106 91 Stockholm, Sweden

Received January 26, 2007; Revised March 19, 2007; Accepted April 17, 2007

ABSTRACT

The *Pcons.net* Meta Server (http://pcons.net) provides improved automated tools for protein structure prediction and analysis using consensus. It essentially implements all the steps necessary to produce a high quality model of a protein. The whole process is fully automated and a potential user only submits the protein sequence. For PSI-BLAST detectable targets, an accurate model is generated within minutes of submission. For more difficult targets the sequence is automatically submitted to publicly available fold-recognition servers that use more advanced approaches to find distant structural homologs. The results from these servers are analyzed and assessed for structural correctness using Pcons and ProQ; and the user is presented with a ranked list of possible models. In addition, if the protein sequence contains more than one domain, these are automatically parsed out and resubmitted to the server as individual queries.

INTRODUCTION

Reliable and accurate predictions of protein structure are important for many biologists. For many years it was believed that manual experts significantly outperformed all automatic methods. However since consensus-based approaches (1) were introduced it has been found that at the most a handful experts in the world can outperform the 'community' of web-servers. It has also been shown consistently in CASP that consensus methods are superior compared to individual methods in predicting the structure of a protein sequence (2–4). Pcons has been among the top performing automated predictors since CASP5 and was the best method for assessing model quality in CASP7 (5).

Here, we introduce the *Pcons.net* meta server (http://pcons.net) which provides improved automated tools for protein structure prediction and analysis using consensus. The whole process is fully automated and a potential user only submits the protein sequence. This makes it easy to acquire structural information without any prior knowledge of remote homology detection, model building and model quality assessment. Pcons has previously been available as a downloadable program as well as through several other meta servers (genesilico.pl and bioinfo.pl). *Pcons.net* meta server provides significant improvements over these servers. It has an improved web interface and prediction accuracy, the local accuracy for each residue is also provided and for easy targets an accurate 3D model is build within minutes of submission.

SERVER DESCRIPTION

The *Pcons.net* Meta Server (http://pcons.net) essentially implements all the steps necessary to produce a high quality model of a protein sequence:

1. Finding the best possible template.
2. Aligning the template to the query sequence.
3. Building a 3D structure based on the alignment.
4. Assessing the quality of the model.

An overview of the method is shown in Figure 1. In the first step domains are assigned using Pfam (6) and a quick database search against known protein structures (PDB90) is performed using BLAST (7) and RPS-BLAST (8). This also establishes the difficulty of the submitted sequence. If a significant hit is found using RPS-BLAST, an all-atom model is produced using, Pfrag, a novel rapid homology modeling program based on segment matching and assembly. If the sequence identity is above 50% this model will be quite close to the native structure, comparable to low-resolution X-ray and NMR structures (9,10). The whole process from sequence to all-atom model takes ~30 s, making it one of the fastest comparative modeling servers available.

RPS-BLAST is also used to parse the sequence into structural domains by analyzing the significance and span of the best RPS-BLAST alignment. If the hit is (i) significant (10^{-5}) and (ii) the alignment contains more than 30 unaligned residues, the unaligned residues are parsed out and resubmitted to the servers as a separate submission. In many cases, these domains agree well with the domains obtained using Pfam.

*To whom correspondence should be addressed. Tel: +1 206 616 4396; Fax: +1 206 685 1792; Email: bjornwa@u.washington.edu

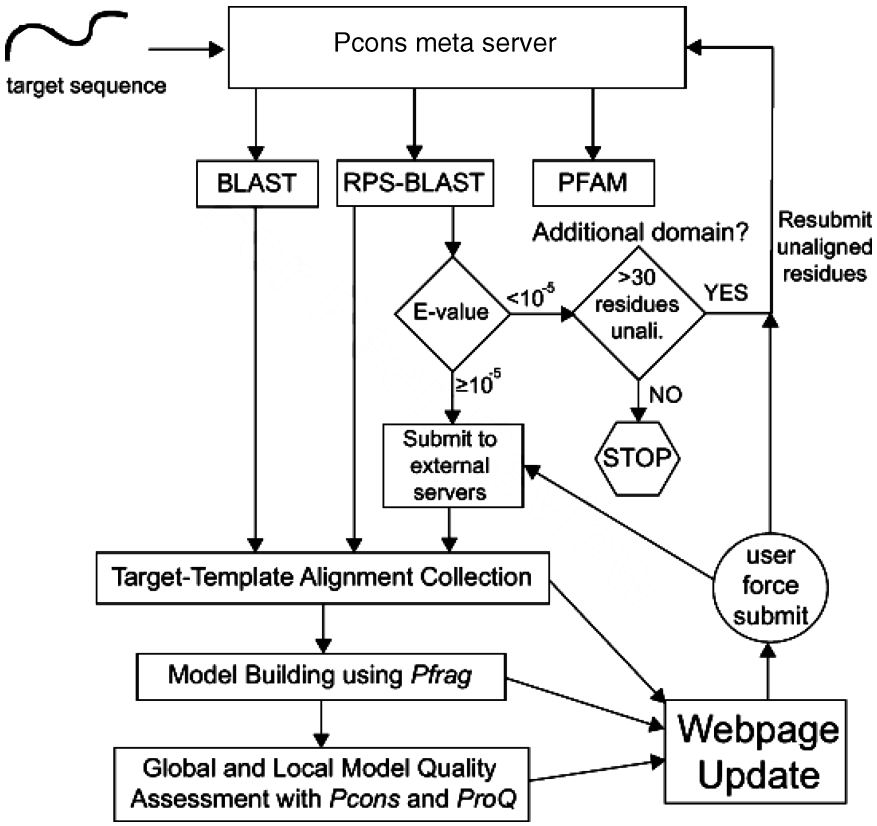

Figure 1. Flow chart describing the different components of *Pcons.net*.

It is only if no significant hits are found using RPS-BLAST, that the sequence is submitted to publicly available more advanced fold-recognition servers (Table 1). The user has the possibility to force the submission of sequences that has clear RPS-BLAST hits. However, we strongly discourage overuse of this possibility in order to not overload the external servers with trivial queries.

The alignments from the initial BLAST, RPS-BLAST as well as the alignments from the fold-recognition servers are collected as they finish and all-atom models are built using Pfrag. When the model building is finished, the quality of the models is assessed using Pcons (1,2,11). Pcons benefits from the use of as many individual servers as possible. Thus, it is important to not put too much weight on a consensus analysis that is only based on the results from a few servers. In parallel to the consensus analysis, the model quality is also assessed purely based on structural features using ProQ (12). Both Pcons and ProQ give an overall quality to each model as well as a local quality score to each individual residue (13). In CASP7, Pcons was one of the best method for assessing the overall quality of protein models and the best method for assessing the local quality of residues (5).

In summary, the major advances over other web servers are:

1. For PSI-BLAST detectable targets a quite accurate homology model is generated within minutes.
2. A query sequence with PSI-BLAST detectable domains is automatically parsed into domains.
3. A novel approach to display alignment similarity makes it easy to quickly select the best model.

Table 1. Internal and external servers utilized by the *Pcons.net* Meta Server. For similar servers, e.g. bas_b and bas_c only one of them is used in the consensus analysis

Servers	URL
BLAST (7)	run internally
RPS-BLAST (8)	run internally
FFAS03 (23)	http://bioinfo.pl/meta/
Meta-Basic (24)	http://bioinfo.pl/meta/
bas_c (24)	http://bioinfo.pl/meta/
bas_b (24)	http://bioinfo.pl/meta/
orfeus2 (25)	http://bioinfo.pl/meta/
SAM-T02 (26)	http://www.cse.ucsc.edu/compbio/HMM-apps/T02-query.html
mGenTHREADER (27)	http://bioinf.cs.ucl.ac.uk/psipred/psiform.html
FUGUE (28)	http://tardis.nibio.go.jp/fugue/
SP^3 (29)	http://sparks.informatics.iupui.edu/hzhou/anonymous-fold-sp3.html
inub (30)	http://inub.cse.buffalo.edu/
FORTE (31)	http://www.cbrc.jp/htbin/forte-cgi/forte_form.pl
HHpred (32)	http://toolkit.tuebingen.mpg.de/hhpred
PSIPRED (18)	http://bioinf.cs.ucl.ac.uk/psipred/psiform.html
Pfam (6)	http://www.sanger.ac.uk/Software/Pfam/

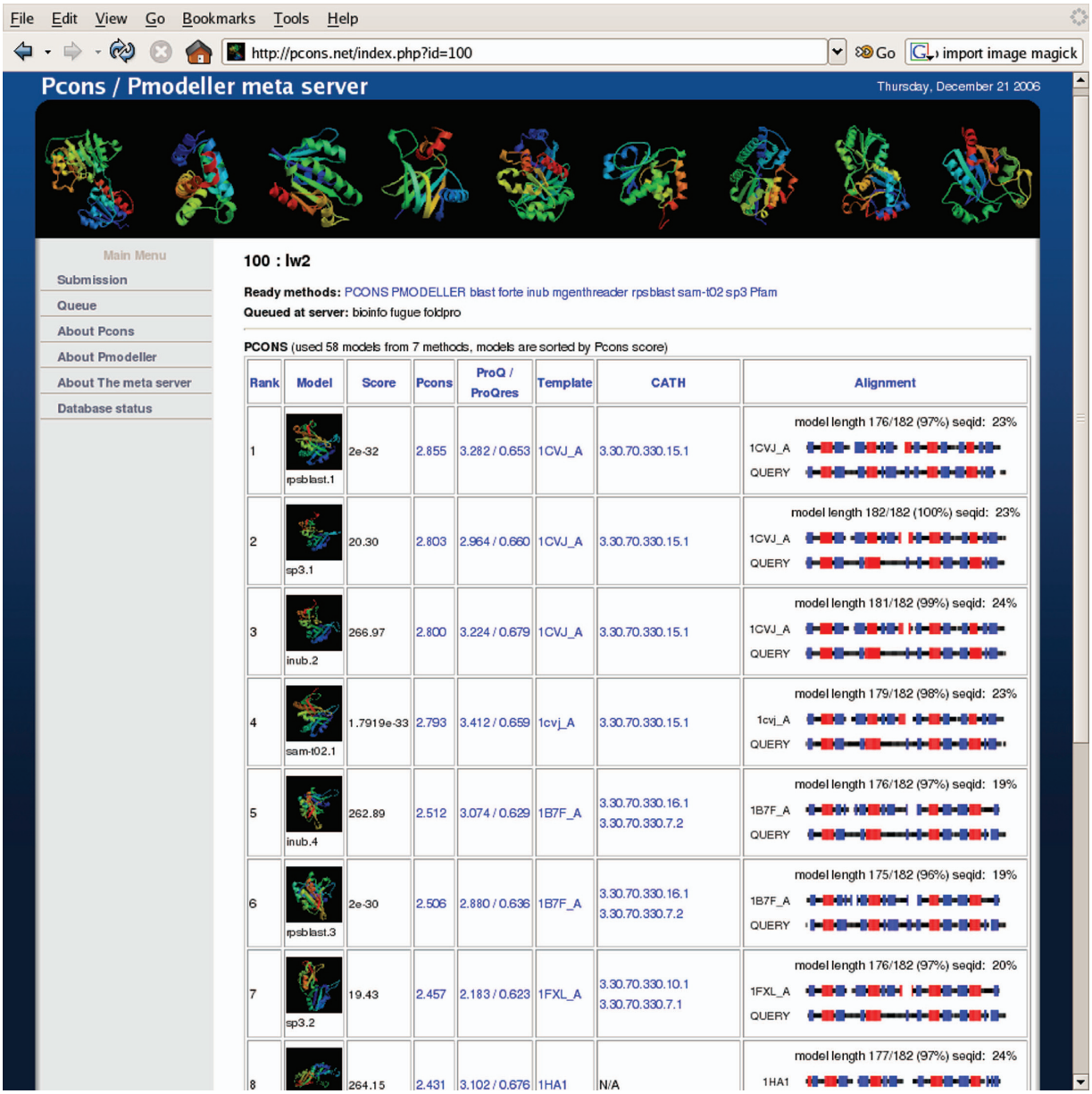

Figure 2. An example of structure prediction results.

4. The overall as well as local quality of the model is assessed, using state-of-the-art methods.

SERVER INPUTS AND OUTPUTS

The server takes a protein sequence in one-letter amino acid format as input. The user has the possibility to name the sequence and to give their e-mail address. Both the name and e-mail address can be used to filter the results in the job queue (http://pcons.net/index.php?queue). Results for a specific job are provided through the web interface by clicking on the job id listed in the job queue table (Figure 2). This page is updated continuously as more predictions are finished. If an e-mail is provided the top 10 ranked model coordinates are e-mailed after 46 h. The 46 h time limit is set to allow for as many fold-recognition servers as possible to finish and provide the basis for the consensus analysis. However, if a significant hit indeed is found using the locally run RPS-BLAST, an accurate model should be ready within minutes of submission.

In addition to the web interface, the *Pcons.net* meta server will also be made available as a web service using the Web Service Description Language (WSDL) (14). The idea behind web services is to allow applications to communicate with each other in a standardized way. WSDL is used to conceptually describe the operations

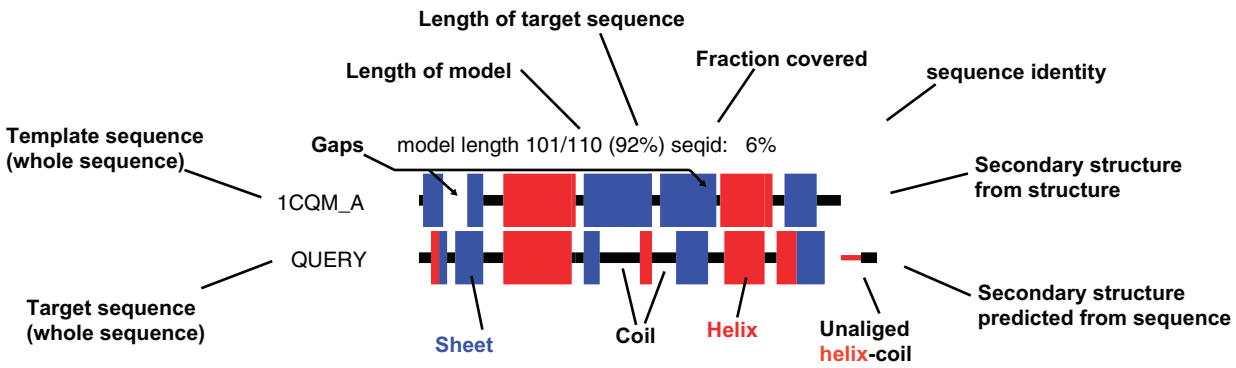

Figure 3. Alignment representation that facilitates comparisons of many different alternative alignments.

available at the service, and expresses the data formats using XML Schema definitions. Communication between web services and clients is done using the SOAP language (Simple Object Application Protocol) (15). For *Pcons.net* this will mean that a user who has access to a web service client, such as Taverna (16), will be able to make submissions to the meta server and also build in these submissions into more complex analysis workflows.

ALIGNMENT REPRESENTATION

An additional novel feature is the representation of the different alignments (Figure 3), which enables a quick overview of the alignment quality and facilitates comparisons of many alternative alignments.

The alignment is represented as a line that is color-coded according to the secondary structure. For the template structure STRIDE (17) is used to assign secondary structure based on the coordinates, for the target sequence PSIPRED (18) is used to predict secondary structure and assign it to each residue. Both the target and the template sequence are represented as full-length sequences, making it possible to see which parts of the target and template that are covered; and if the alignment spans only a part of the whole template structure.

Here, the user also has the possibility to submit unaligned regions that did not fulfill the criteria for automatic domain resubmission (see above).

MODEL BUILDING

The model building based on the target–template alignment is performed using Pfrag, a reimplementation of the SegMod (19) homology modeling program. It builds models based on segment matching. By searching a database of highly refined protein structures, structural fragments are found that matches the template structure as closely as possible. Criteria for evaluating individual fragments are the degree of amino acid sequence homology between the target and the template, the RMSD deviation between a fragment and the template structure and the Lennard–Jones interaction energy between fragments and the structure. Initial screening of fragments is done using the methodology of distance matching by Jones and Thirup (20). The all-atom models are then energy minimized using the ENCAD force field (21) to enforce proper stereochemistry.

QUALITY SCORES

A key component for any successful protein structure protocol is the ability to assign quality scores to the created models. *Pcons.net* scores models using the best methods currently available. For each model three global quality scores are provided, one based on consensus (Pcons), one based solely on structure (ProQ) and one using a combination of the two (Pmodeller). All are presented in the job summary page. The reason for providing more than one score is that they contain complementary information. The Pcons score, for instance, is only meaningful if a sufficient number of models are available. If this is not the case, a structural evaluation using ProQ might be more suitable and for other cases the ProQ score might be a useful aid in the process of choosing the best model.

From a user perspective it is important to know when to trust a certain score. Based on results from the quality assessment category in CASP7 (5) the Pcons score correlates well with the correct quality of the models as measured by LGscore (22) ($R = 0.96$). Moreover a Pcons score above 1.1 separates correct from incorrect models almost perfectly (only 2.5% false predictions). The ProQ and Pmodeller scores are the predicted LGscore and score values above 1.5 correspond to *P*-values better than 10^{-3}.

In addition to the global quality scores, each amino acid in the models is given an estimate of the CA–CA error as measured by the local S-score ($S = 1/(1 + \text{error}^2/5)$). The S-score varies between 0 and 1 corresponding to high and low error, respectively, e.g. if the S-score is larger than 0.5 the error is predicted to be <2.24 Å ($5^{1/2}$). The advantage with this type of score is that it focusses on the regions that have low error and gives the same score value for regions that are wrong. As for the global scores the local quality is predicted using either consensus (Pcons) or structural features (ProQres). In terms of performance, Pcons is superior to ProQres (13). In fact, no non-consensus-based approach is nearly as good as

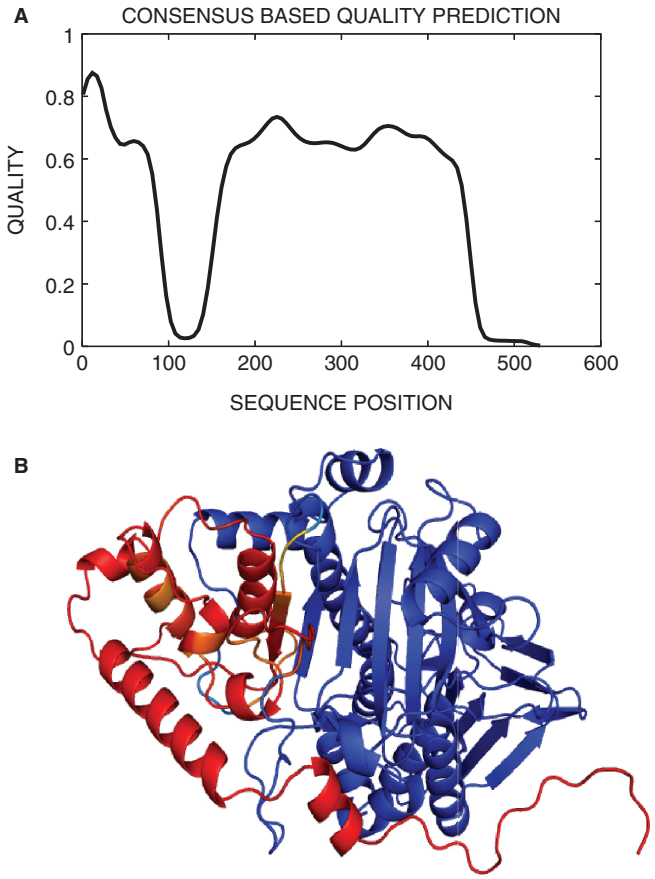

Figure 4. Local quality prediction using Pcons. (**A**) Predicted quality plotted for each residue in the sequence. (**B**) The structure color-coded from red to blue using the predicted quality, corresponding to poor and good, respectively (picture made using PyMOL (33). In this particular example, Pcons has identified a region around residue number 100 and the C-terminal to be incorrect. Despite that these two regions are far apart in sequence they end up on the same side of the protein, since the rest of the protein is correct; this suggests that the C-terminal residues makes some interactions with residues in other region that is not capture by this model. With this information it might be possible to improve the model.

consensus-based approaches (5). However, ProQres still provide some additional value as a complement when there is no clear consensus or as additional augmentation when the consensus is weak. The local quality predictions are accessible by clicking either on the Pcons score or on the ProQ score in the job summary page (Figure 2). The local quality scores predicted by Pcons are also added to the B-factor column of all models for easy visualization in any coordinate viewing program (Figure 4).

THROUGHPUT

The throughput of *Pcons.net* depends to a large degree on the difficulty of the target. For the easy targets, the meta server could easily handle more than 1000 requests per day. But for the harder targets it can only handle about 50 requests per day, due to the throughput of the external server it uses. To avoid overloading the external servers there is also a limit in the number of pending external server jobs the meta server can have. If this limit is reached, the meta server will queue the jobs locally until the number of pending jobs decreases.

ACKNOWLEDGEMENTS

First of all we want to thank all developers of servers. Without these the consensus approach would not have any value. The success of consensus-based methods should really be attributed to the whole collective force of fold-recognition method developers and we encourage users of Pcons.net to cite the individual servers as well. We would also like to thank Michael Levitt for kindly providing the source code to SegMod and Erik Lindahl for scientific advise.

This work was supported by grants from the Swedish Research Councils and the EU 6th Framework Program is gratefully acknowledged for support to the GeneFun project, contract LSHG-CT-2004-503567 and to the EMBRACE project, contract LHSG-CT-2004-512092. Funding to pay the Open Access publication charges for this article was provided by the EMBRACE project

.

Conflict of interest statement. None declared.

REFERENCES

1. Lundström,J., Rychlewski,L., Bujnicki,J. and Elofsson,A. (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci.*, **10**, 2354–2362.
2. Wallner,B., Fang,H. and Elofsson,A. (2003) Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. *Proteins*, **53**(Suppl. 6), 534–541.
3. Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2003) Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*, **53**(Suppl. 6), 334–339.
4. Kryshtafovych,A., Venclovas,C., Fidelis,K. and Moult,J. (2005) Progress over the first decade of CASP experiments. *Proteins*, **61**(Suppl. 7), 225–236.
5. Wallner,B. and E. Elofsson,A. (2007) Assessment of global and local quality model in casp7 using pcons. *Manuscript in preparation*.
6. Sonnhammer,E., Eddy,S. and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.
7. Altschul,S., Madden,T., Schaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
8. Marchler-Bauer,A., Panchenko,A.R., Shoemaker,B.A., Thiessen,P.A., Geer,L.Y. and Bryant,S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 281–283.
9. Marti-Renom,M., Stuart,A., Fiser,A., Sánchez,R., Melo,F. and Sali,A. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291–325.
10. Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.
11. Wallner,B. and Elofsson,A. (2005) All are not equal: a benchmark of different homology modeling programs. *Protein Sci.*, **14**, 1315–1327.
12. Wallner,B. and Elofsson,A. (2003) Can correct protein models be identified? *Protein Sci.*, **12**, 1073–1086.

13. Wallner,B. and Elofsson,A. (2006) Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Sci.*, **15**, 900–913.
14. Web services description language. http://www.w3.org/TR/wsdl
15. Simple object access protocol. http://www.w3.org/TR/soap
16. Hull,D., Wolstencroft,K., Stevens,R., Goble,C., Pocock, M.R.,Li, P. and Oinn,T. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflow. *Bioinformatics*, **20**, 3045–3054.
17. Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.
18. Jones,D. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
19. Levitt,M. (1992) Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.*, **226**, 507–533.
20. Jones,T. A. and Thirup,S. (1986) Using known substructures in protein model building and crystallography. *EMBO J.*, **5**, 819–822.
21. Levitt,M. (1983) Molecular dynamics of native protein. i. computer simulation of trajectories. *J. Mol. Biol.*, **168**, 595–617.
22. Cristobal,S., Zemla,A., Fischer,D., Rychlewski,L. and Elofsson,A. (2001) A study of quality measures for protein threading models. *BMC Bioinformatics*, **2**(5).
23. Jaroszewski,L., Rychlewski,L., Li,Z., Li,W. and Godzik,A. (2005) FFAS03: a server for profile-profile sequence alignments. *Nucleic Acids Res.*, **33**(Web Server issue), W284–W288.
24. Ginalski,K., von Grotthuss,M., Grishin,N. V. and Rychlewski,L. (2004) Detecting distant homology with meta-BASIC. *Nucleic Acids Res.*, **32**(Web Server issue), W576–W581.
25. Ginalski,K., Pas,J., Wyrwicz,L. S., von Grotthuss,M., Bujnicki,J. M. and Rychlewski,L. (2003) ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acids Res.*, **31**, 3804–3807.
26. Karplus,K., Karchin,R., Draper,J., Casper,J., Mandel-Gutfreund,Y., Diekhans,M. and Hughey,R. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction.. *Proteins*, **53**(Suppl. 6), 491–496.
27. McGuffin,L. J. and Jones,D. T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, **19**, 874–881.
28. Shi,J., Blundell,T. and Mizuguchi,K. (2001) Fugue: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**, 243–257.
29. Zhou,H. and Zhou,Y. (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins*, **58**, 321–328.
30. Fischer,D. (2003) 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. *Proteins*, **51**, 434–441.
31. Tomii,K. and Akiyama,Y. (2004) FORTE: a profile-profile comparison tool for protein fold recognition. *Bioinformatics*, **20**, 594–595.
32. Soding,J., Biegert,A. and Lupas,A. N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**(Web Server issue), W244–W248.
33. DeLano,W. (2002) The pymol molecular graphics system. http://www.pymol.org