Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W407–W410***
doi:10.1093/nar/gkm290

ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins

Markus Wiederstein and Manfred J. Sippl*

Center of Applied Molecular Engineering, Division of Bioinformatics, University of Salzburg, Hellbrunnerstrasse 34, 5020 Salzburg, Austria

Received January 31, 2007; Revised March 30, 2007; Accepted April 12, 2007

ABSTRACT

A major problem in structural biology is the recognition of errors in experimental and theoretical models of protein structures. The ProSA program (Protein Structure Analysis) is an established tool which has a large user base and is frequently employed in the refinement and validation of experimental protein structures and in structure prediction and modeling. The analysis of protein structures is generally a difficult and cumbersome exercise. The new service presented here is a straightforward and easy to use extension of the classic ProSA program which exploits the advantages of interactive web-based applications for the display of scores and energy plots that highlight potential problems spotted in protein structures. In particular, the quality scores of a protein are displayed in the context of all known protein structures and problematic parts of a structure are shown and highlighted in a 3D molecule viewer. The service specifically addresses the needs encountered in the validation of protein structures obtained from X-ray analysis, NMR spectroscopy and theoretical calculations. ProSA-web is accessible at https://prosa.services.came.sbg.ac.at

INTRODUCTION

The availability of a structural model of a protein is one of the keys for understanding biological processes at a molecular level. The recent advances in experimental technology have led to the emergence of large-scale structure determination pipelines aimed at the rapid characterization of protein structures. The resulting amount of experimental structural information is enormous. The application of computational methods for the prediction of unknown structures adds another plethora of structural models. The latest NAR web server issue, e.g. lists about 50 tools in the category '3D Structure Prediction' (1). The assessment of the accuracy and reliability of experimental and theoretical models of protein structures is a necessary task that needs to be addressed regularly and in particular, it is essential for maintaining integrity, consistency and reliability of public structure repositories (2).

ProSA (3) is a tool widely used to check 3D models of protein structures for potential errors. Its range of application includes error recognition in experimentally determined structures (4–6), theoretical models (7–10) and protein engineering (11,12). Here we present a web-based version of ProSA, ProSA-web, that encompasses the basic functionality of stand-alone ProSA and extends it with new features that facilitate interpretation of the results obtained. The overall quality score calculated by ProSA for a specific input structure is displayed in a plot that shows the scores of all experimentally determined protein chains currently available in the Protein Data Bank (PDB) (13). This feature relates the score of a specific model to the scores computed from all experimental structures deposited in PDB. Problematic parts of a model are identified by a plot of local quality scores and the same scores are mapped on a display of the 3D structure using color codes.

A particular intention of the ProSA-web application is to encourage structure depositors to validate their structures before they are submitted to PDB and to use the tool in early stages of structure determination and refinement. The service requires only C^{α} atoms so that low-resolution structures and approximate models obtained early in the structure determination process can be evaluated and compared against high-resolution structures. The ProSA-web service returns results instantaneously, i.e. the response time is in the order of seconds, even for large molecules.

*To whom correspondence should be addressed. Tel: +43-662-8044-5796; Fax: +43-662-8044-176; Email: sippl@came.sbg.ac.at

WEB SERVER USAGE

Required input

ProSA-web requires the atomic coordinates of the model to be evaluated. Users can supply coordinates either by uploading a file in PDB format or by entering the four-letter code of a protein structure available from PDB. A chain identifier and an NMR model number may be used to specify a particular model. A list with possible values of these parameters is presented to the user if the entered chain identifier or model number is invalid. If no chain identifier or model number is supplied by the user, the first chain of the first model found in the PDB file is used for analysis.

Range of computations

The computational engine used for the calculation of scores and plots is standard ProSA which uses knowledge-based potentials of mean force to evaluate model accuracy (3). All calculations are carried out with C^{α} potentials, hence ProSA-web can also be applied to low-resolution structures or other cases where the C^{α} trace is available only (a set of C^{β} potentials is included in the stand-alone version of ProSA, see Supplementary Data 1). After parsing the coordinates, the energy of the structure is evaluated using a distance-based pair potential (14,15) and a potential that captures the solvent exposure of protein residues (16). From these energies, two characteristics of the input structure are derived and displayed on the web page: its *z*-score and a plot of its residue energies.

The *z*-score indicates overall model quality and measures the deviation of the total energy of the structure with respect to an energy distribution derived from random conformations (3,15). *Z*-scores outside a range characteristic for native proteins indicate erroneous structures. In order to facilitate interpretation of the *z*-score of the specified protein, its particular value is displayed in a plot that contains the *z*-scores of all experimentally determined protein chains in current PDB (an example is shown in Figure 1A). Groups of structures from different sources (X-ray, NMR) are distinguished by different colors. This plot can be used to check whether the *z*-score of the protein in question is within the range of scores typically found for proteins of similar size belonging to one of these groups.

The energy plot shows the local model quality by plotting energies as a function of amino acid sequence position *i* (see Figure 1B and D for example). In general, positive values correspond to problematic or erroneous parts of a model. A plot of single residue energies usually contains large fluctuations and is of limited value for model evaluation. Hence the plot is smoothed by calculating the average energy over each 40-residue fragment $s_{i,i+39}$, which is then assigned to the 'central' residue of the fragment at position $i+19$.

In order to further narrow down those regions in the model that contribute to a bad overall score, ProSA-web visualizes the 3D structure of the protein using the molecule viewer Jmol (http://www.jmol.org). Residues with unusually high energies stand out by color from the rest of the structure (Figure 1C and E). The interactive facilities provided by Jmol, like distance measurements, etc. are available for exploring these regions in more detail.

Protein structure validation by example

In what follows, we provide a typical example for the application of ProSA-web in the validation of protein structures. We analyze two structures determined by X-ray analysis and deposited in PDB . The first is the structure of MsbA from *Escherichia coli*, a homolog of the multi-drug resistance ATP-binding cassette (ABC) transporters (PDB code 1JSQ, release date 12 September 2001) determined to a resolution of 4.5 Å (17). The structure consists of an N-terminal transmembrane domain and a soluble nucleotide-binding domain. Doubts regarding the quality of 1JSQ were raised after the X-ray structure of a close homolog became available which turned out to be surprisingly different. This second structure, multi-drug ABC transporter Sav1866 from *Staphylococcus aureus* (PDB code 2HYD, release date 5 September 2006) was determined to a resolution of 3.0 Å (18). Based on the newly determined structure, it was realized that the published structure of the MsbA model is incorrect and as a consequence the related publication had to be retracted (19).

Here, we apply the ProSA-web service to the analysis of the incorrect 1JSQ and the recently released 2HYD model. An interesting aspect is that both structures contain a transmembrane domain. Since the energy functions used in ProSA are derived mainly from soluble globular proteins of known structure, it is not clear in advance to what extent the ProSA scores reflect problems in protein structures containing membrane spanning domains.

Figure 1A–C shows the results of ProSA-web obtained for 1JSQ (chain A). The *z*-score of this model is −0.60, a value far too high for a typical native structure. This can clearly be seen when the score is compared to the scores of other experimentally determined protein structures of the size of 1JSQ (Figure 1A). Furthermore, large parts of the energy plot show highly positive energy values, especially the N-terminal half of the sequence which contains part of the membrane spanning domain (Figure 1B). In the C^{α} trace of the model, residues with high energies are shown in grades of red (Figures 1C), and it is evident from these figures that the N-terminal transmembrane domain as well as the C-terminal globular domain contain regions of offending energies.

Figure 1A also shows the location of the *z*-score for 2HYD (chain A). The value, −8.29, is in the range of native conformations. Overall the residue energies are largely negative with the exception of some peaks in the N-terminal part (Figure 1D). These peaks are supposed to correspond to membrane spanning regions of the protein. In the C^{α} trace, these regions show up as clusters of residues colored in red (Figure 1E, lower left). The C-terminal domain shows a high number of residues colored in blue and an energy distribution that is entirely below the zero base line, consistent with the parameters of a typical protein (Figure 1D and E).

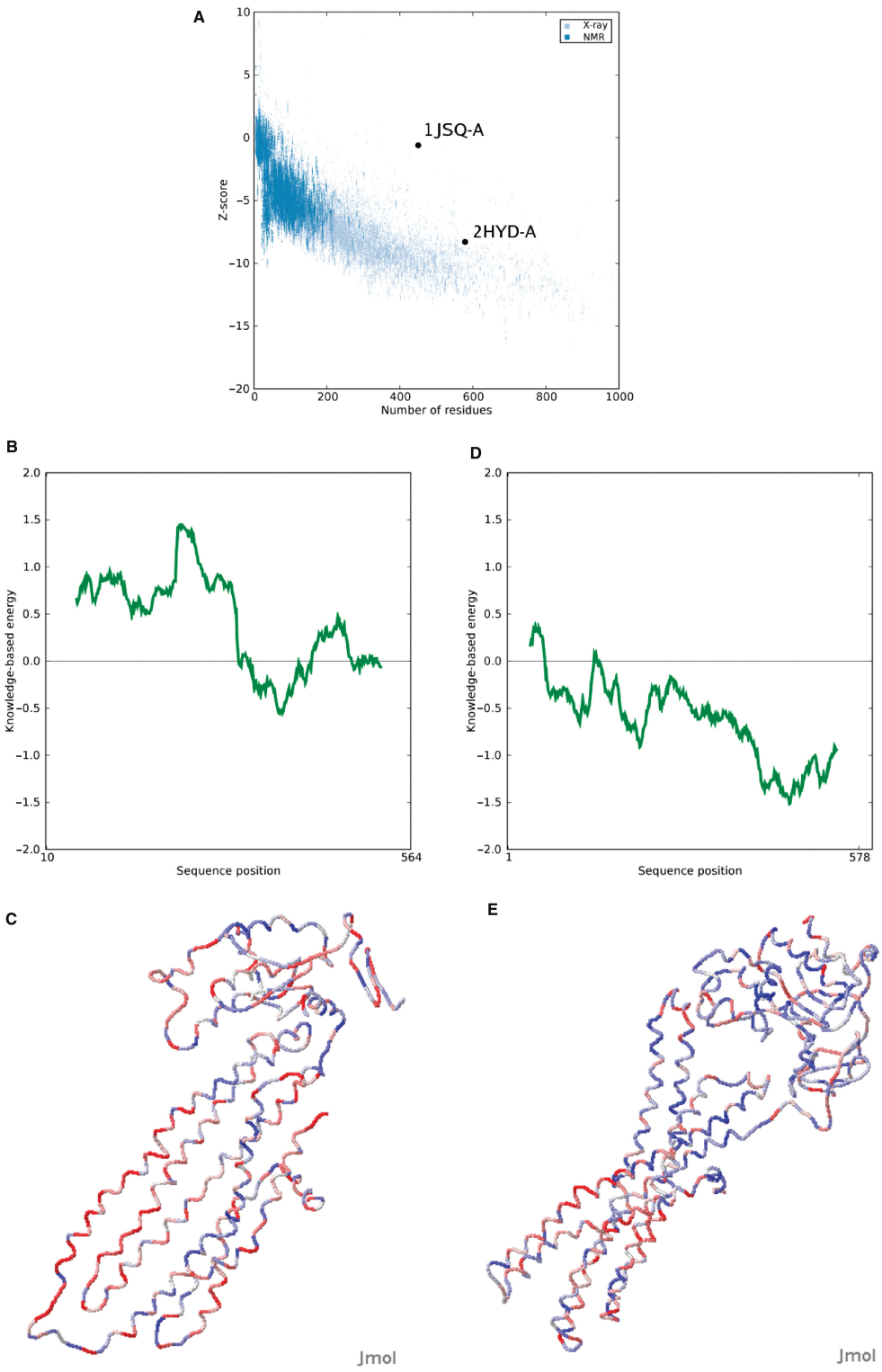

Figure 1. Investigation of two ABC transporter structures using the ProSA-web service. Subfigures (A–C) show the results for a monomer of MsbA (PDB code 1JSQ, chain A (17)). The structure was determined by X-ray crystallography to 4.5 Å resolution and had to be retracted due to problems in the interpretation of the crystallographic raw data (19). Subfigures (A, D and E) show the results for a monomer of Sav1866 (PDB code 2HYD, chain A (18)) as determined by X-ray crystallography to 3.0 Å resolution. Although homologous to 1JSQ, this structure differs considerably from the 1JSQ A chain. The ProSA-web results indicate that 2HYD has features characteristic for native structures. (A) ProSA-web *z*-scores of all protein chains in PDB determined by X-ray crystallography (light blue) or NMR spectroscopy (dark blue) with respect to their length. The plot shows only chains with less than 1000 residues and a *z*-score ≤ 10. The *z*-scores of 1JSQ-A and 2HYD-A are highlighted as large dots. (B) Energy plot of 1JSQ-A. Residue energies averaged over a sliding window are plotted as a function of the central residue in the window. A window size of 80 is used due to the large size of the protein chain (default: 40). (C) Jmol C^{α} trace of 1JSQ-A. Residues are colored from blue to red in the order of increasing residue energy. (D–E) Same as (B–C) but for 2HYD-A.

CONCLUSION

The protein structure community is, to some extent, aware of the fact that the RCSB protein data base contains erroneous structures. But it is quite difficult to spot these errors. Grossly misfolded structures are sometimes revealed after the results of subsequent independent structure determinations become available. Errors in regular PDB files generally remain unknown to the structural community until the corresponding revisions are made available. Hence, diagnostic tools that reveal unusual structures and problematic parts of a structure in a manner that is independent of the experimental data and the specific method employed are essential in many areas of protein structure research.

ProSA is a diagnostic tool that is based on the statistical analysis of all available protein structures. The potentials of mean force compiled from the data base provide a statistical average over the known structures. Structures of soluble globular proteins whose *z*-scores deviate strongly from the data base average are unusual and frequently such structures turn out to be erroneous. For proteins containing membrane spanning regions, the significance of deviations from the average over the data base is less clear.

Here, we provide an example of a published structure (1JSQ) that is known to be incorrect as is revealed by subsequent independent X-ray analysis of a related protein yielding a completely different conformation. The ProSA-web result obtained for 1JSQ shows extreme deviations when compared to all the structures in PDB (Figure 1A). In contrast, the score obtained for the related 2HYD structure is close to the data base average. The result demonstrates that also for membrane proteins large deviations from normality may indicate an erroneous structure.

SUPPLEMENTARY DATA

(1) ProSA stand-alone version: http://cms.came.sbg.ac.at/typo3/index.php?id=prosa_download
(2) List of studies that use ProSA for model validation: http://www.came.sbg.ac.at/typo3/index.php?id=prosa_literature

ACKNOWLEDGEMENTS

The authors are grateful to Christian X. Weichenberger who suggested the use of the ABC transporter structures as an example. This work was supported by FWF Austria, grant number P13710-MOB. Use of the ProSA-II program on the ProSA-web server is granted under an academic license agreement by Proceryon Science for Life GmbH (http://www.proceryon.com) which is gratefully acknowledged. Funding to pay the Open Access publication charges for this article was provided by the University of Salzburg, Austria.

Conflict of interest statement. None declared

REFERENCES

1. Fox,J.A., McMillan,S. and Ouellette,B.F.F. (2006) A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucleic Acids Res.*, **34**, W3–W5.
2. Berman,H.M., Burley,S.K., Chiu,W., Sali,A., Adzhubei,A., Bourne,P.E., Bryant,S.H., Dunbrack,R.L., Fidelis,K. *et al.* (2006) Outcome of a workshop on archiving structural models of biological macromolecules. *Structure*, **14**, 1211–1217.
3. Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355–362.
4. Banci,L., Bertini,I., Cantini,F., DellaMalva,N., Herrmann,T., Rosato,A. and Wüthrich,K. (2006) Solution structure and intermolecular interactions of the third metal-binding domain of ATP7A, the Menkes disease protein. *J. Biol. Chem.*, **281**, 29141–29147.
5. Llorca,O., Betti,M., Gonzlez,J.M., Valencia,A., Mrquez,A.J. and Valpuesta,J.M. (2006) The three-dimensional structure of an eukaryotic glutamine synthetase: functional implications of its oligomeric structure. *J. Struct. Biol.*, **156**, 469–479.
6. Teilum,K., Hoch,J.C., Goffin,V., Kinet,S., Martial,J.A. and Kragelund,B.B. (2005) Solution structure of human prolactin. *J. Mol. Biol.*, **351**, 810–823.
7. Petrey,D. and Honig,B. (2005) Protein structure prediction: inroads to biology. *Mol. Cell*, **20**, 811–819.
8. Ginalski, K. (2006) Comparative modeling for protein structure prediction. *Curr. Opin. Struct. Biol.*, **16**, 172–177.
9. Panteri,R., Paiardini,A. and Keller,F. (2006) A 3D model of Reelin subrepeat regions predicts Reelin binding to carbohydrates. *Brain Res.*, **1116**, 222–230.
10. Mansfeld,J., Gebauer,S., Dathe,K. and Ulbrich-Hofmann,R. (2006) Secretory phospholipase A2 from Arabidopsis thaliana: insights into the three-dimensional structure and the amino acids involved in catalysis. *Biochemistry*, **45**, 5687–5694.
11. Beissenhirtz,M.K., Scheller,F.W., Viezzoli,M.S. and Lisdat,F. (2006) Engineered superoxide dismutase monomers for superoxide biosensor applications. *Anal. Chem.*, **78**, 928–935.
12. Wiederstein,M. and Sippl,M.J. (2005) Protein sequence randomization: efficient estimation of protein stability using knowledge-based potentials. *J. Mol. Biol.*, **345**, 1199–1212.
13. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
14. Sippl,M.J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, **213**, 859–883.
15. Sippl,M.J. (1995) Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.*, **5**, 229–235.
16. Sippl,M.J. (1993) Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput. Aided Mol. Des.*, **7**, 473–501.
17. Chang,G. and Roth,C.B. (2001) Structure of MsbA from E. coli: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters. *Science*, **293**, 1793–1800.
18. Dawson,R.J.P. and Locher,K.P. (2006) Structure of a bacterial multidrug ABC transporter. *Nature*, **443**, 180–185.
19. Chang,G., Roth,C.B., Reyes,C.L., Pornillos,O., Chen,Y.-J. and Chen,A.P. (2006) Retraction. *Science*, **314**, 1875.