Nucleic Acids Research, 2006, Vol. 34, Web Server issue W231–W234
doi:10.1093/nar/gkl314

HARMONY: a server for the assessment of protein structures

G. Pugalenthi, K. Shameer, N. Srinivasan[1] and R. Sowdhamini*

National Centre for Biological Sciences (TIFR), GKVK Campus, Bellary Road Bangalore 560 065, India and [1]Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

Received February 9, 2006; Revised March 1, 2006; Accepted April 11, 2006

ABSTRACT

Protein structure validation is an important step in computational modeling and structure determination. Stereochemical assessment of protein structures examine internal parameters such as bond lengths and Ramachandran (ϕ,ψ) angles. Gross structure prediction methods such as inverse folding procedure and structure determination especially at low resolution can sometimes give rise to models that are incorrect due to assignment of misfolds or mistracing of electron density maps. Such errors are not reflected as strain in internal parameters. HARMONY is a procedure that examines the compatibility between the sequence and the structure of a protein by assigning scores to individual residues and their amino acid exchange patterns after considering their local environments. Local environments are described by the backbone conformation, solvent accessibility and hydrogen bonding patterns. We are now providing HARMONY through a web server such that users can submit their protein structure files and, if required, the alignment of homologous sequences. Scores are mapped on the structure for subsequent examination that is useful to also recognize regions of possible local errors in protein structures. HARMONY server is located at http://caps.ncbs.res.in/harmony/

INTRODUCTION

Knowledge of the three-dimensional (3D) structure of a protein is essential to understand its function. Protein structures are determined by experimental methods, such as X-ray crystallography and nuclear magnetic resonance (NMR), or modeled by comparative (1–3) or combinatorial modeling techniques (4,5) or by *de novo* methods (6,7) and deposited in the Protein Data Bank (PDB) (8). When the 3D structure of a protein is determined by X-ray analysis, local errors can arise as a consequence of poorly defined electron density, particularly because of flexibility or disorder in loops. Global errors can occur occasionally because of misinterpretation or mistracing of the polypeptide chain. In comparative modeling, the quality of a model is highly dependent on correct template selection and sequence alignment between query and template (9). Combinatorial modeling or fold prediction methods can also lead to incorrect models. Such incorrect models are not easily identified by checking internal parameters or by calculating energy values.

Various methods have been developed for protein structure validation and a majority of the current servers examine internal parameters, strains in bond torsions, disallowed conformations or by calculating the energies. These approaches mostly assess the stereochemical quality of the protein structures and provide no indication for the compatibility of the sequence to the structure. Few approaches assess the sequence–structure compatibility by analyzing the propensity of amino acid residues to be present in particular local environments (10–12) or psuedoenergies of pairwise residue interactions (13) or other similar parameters (14–17). Most of these approaches can perform well in the prediction of global errors in the gross structure. A few methods, such as those developed by Sippl *et al.* (Prosa) (13) and Eisenberg *et al.* (10) attempt to recognize errors in the local regions of protein structures with grossly correct fold. In this article, we report the availability of a web-based algorithm which employs the frequency of an amino acid type (propensity) as well as the frequency of amino acid replacements (substitution) in a particular structural environment to discriminate gross as well as local misfolds in protein structures. The results of our approach have been compared with other relevant methods (as in http://xray.bmc.uu.se/gerard/embo2001/modval/links.html) and have been provided in Supplementary Data.

Brief outline of HARMONY algorithm

The approach follows the procedure outlined by Overington *et al.* (18) and Topham *et al.* (11). Structural environment

*To whom correspondence should be addressed. Tel: +91 80 23636421, ext. 4240; Fax: +91 80 23636462; Email: mini@ncbs.res.in

of an amino acid is described by its backbone conformation (nine types), hydrogen bonding (two types) and solvent accessibility (three types) patterns following JOY representation (18,19). Raw scores of amino acid occurrences and exchanges are considered after a suitable normalization (20) and arranged as amino acid propensity table and 54 amino acid substitution tables, respectively. Sequence homologues for each of the query sequences can be identified from the Swiss-Prot database (21) or the non-redundant (NR) protein sequence database using BLAST (22). Homologous proteins with no more than 90% sequence identity with one another were aligned with the query using MALIGN (23). Propensity and substitution values derived for the query and that of large number of unrelated globular domains are compared by χ^2 test and further smoothened by a 11-residue window to recognize possible local errors in the local regions. The total score for the query protein structure is useful in detecting gross errors when viewed along with calibrated entries. However, for the detection of local errors, the smoothened scores over a moving window of contiguous residues were compared with those of a random sequence of same amino acid composition as the query. We employ the reversed query sequence as a reference for the detection of local errors.

HARMONY server: description and features

The HARMONY server can be basically divided into three subsystems: (i) the web interface system, which is written with HTML and PERL CGI (ii) the background process system, which is written with FORTRAN77 and PERL and (iii) a reporting system for graphical and textual output through dynamic PERL CGI programs (Figure 1). The web interface subsystem mainly deals with receiving information from the user and checking the validity of the submitted data. The background processing subsystem performs all the computation of validation like extracting homologues from the sequence database, aligning the homologous sequence with query sequence, calculating substitution and propensity score at each residue position and validating the structure.

Input

In order to assess the quality of query protein, the user may upload query structure in the PDB format. Options are provided to obtain homologues for the query from various sequence databases (the Swiss-Prot database and the NR protein sequence database) using the BLAST program at different levels of stringency through expectation values. Alternately, the user can also upload both the structure and curated

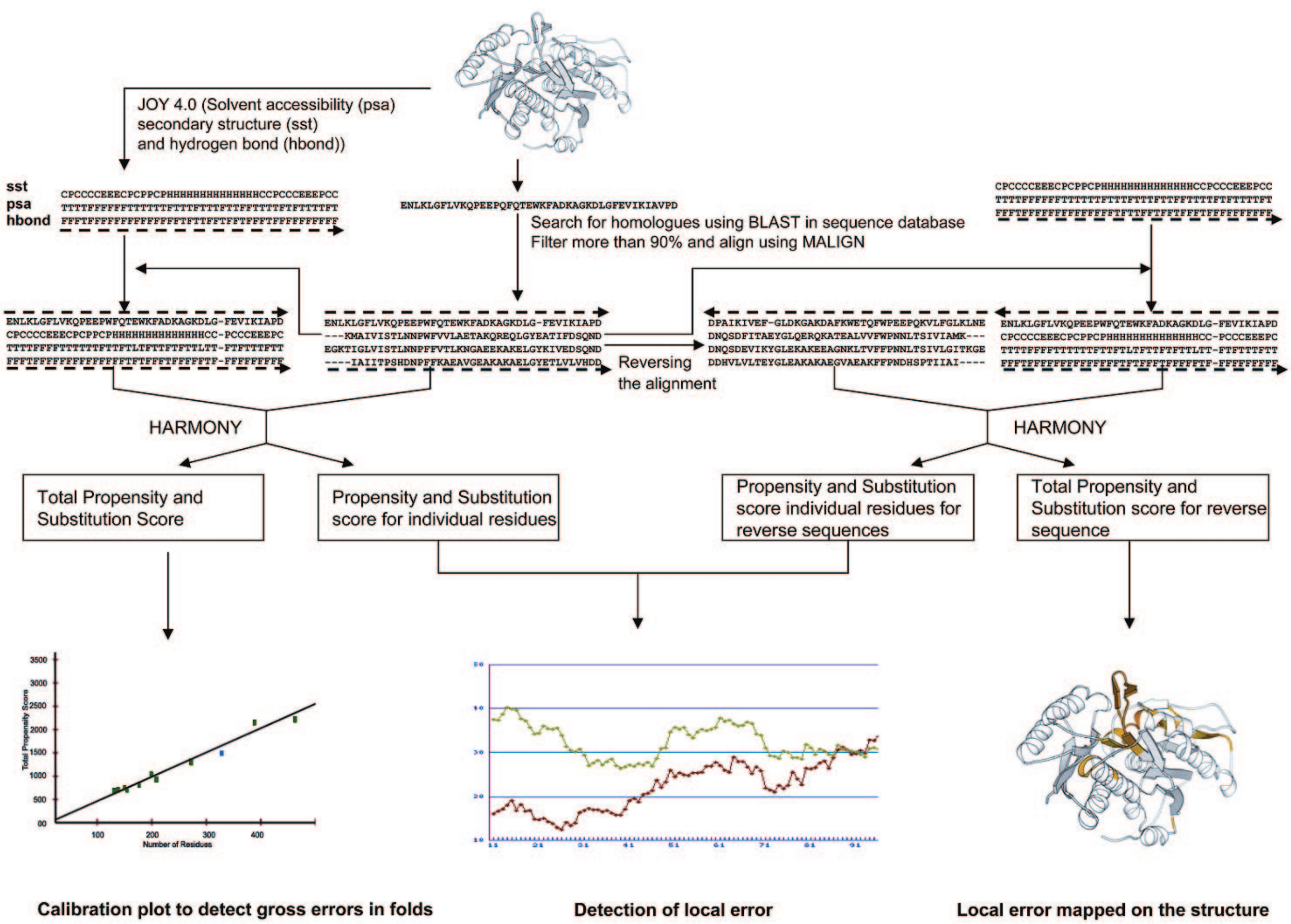

Figure 1. Flowchart representing the steps involved in structure validation for a given query.

multiple sequence alignment of homologous sequences. The length of the window can be chosen from a dropdown menu by the user. Pre-processing involves check and correction of any mislabeled atoms and alternate occupancy.

Output

- Misfolds and folds with elaborate regions of errors can be identified by plotting the scores of the query protein on a calibration plot using known structures. Proteins with misfold regions attain low scores and appear well below the straight fit line.
- The raw and smoothened values of propensity and substitution scores for each residue are provided in tabular format.
- Further, a graph representation provides the smoothened scores between query sequence in comparison with the reverse sequence. The reverse sequence and its scores are used as a control to identify local errors in the proposed protein model. Regions where the reverse sequence acquires better substitution score indicate possible local errors.
- Difference between the scores of the actual and reverse sequence is mapped on the structure colored according to the degree of the local error. Static image of the 3D structure is provided using MOLSCRIPT (24). The structure of the query protein with regions colored using HARMONY scores can be downloaded and conveniently displayed with structural viewers such as RASMOL (25).
- HARMONY server run on a single structure takes only 30 s when homologous sequences are searched internally against Swiss-Prot database

Benchmarking study

HARMONY has been tested on 4020 protein domains of varying length (Supplementary Data) (26). A direct correlation is expected between HARMONY scores and protein size for folds that are highly compatible with the sequence. In this dataset, a high correlation co-efficient was obtained including small domains and disulphide-rich folds. Gross misfolds can be identified as models with significantly lower scores (Figure 2a). We also performed a limited Novotny-type test on immunoglobulin sequence (PDB code: 1mcp) by threading the hemerythrin structure. This deliberate misfold obtained low score since the environments of residues in the incorrect 3D structures are not compatible with the residues in the corresponding positions of the sequence (Figure 2b and c). We have further compared the results of our approach with other methods (27–30) (see Supplementary Data). For instance, for the crystal structure of *Escherichia coli* single-stranded DNA-binding protein (PDB code: 1qvc) solved at resolution 2.2 Å (31), chain A of 1qvc has an apparent sequence–structure mapping error which is caused by a one-residue register shift in position 100 and 110 (32). On a test dataset of 13 error-prone proteins (32), the current method performs equal to or better than the other methods. HARMONY server detects local error in the region consisting of residues 103–140 (Figure 2d).

CONCLUSIONS

Errors in the overall fold or in the local regions can be reflected as poor network of hydrogen bonds or

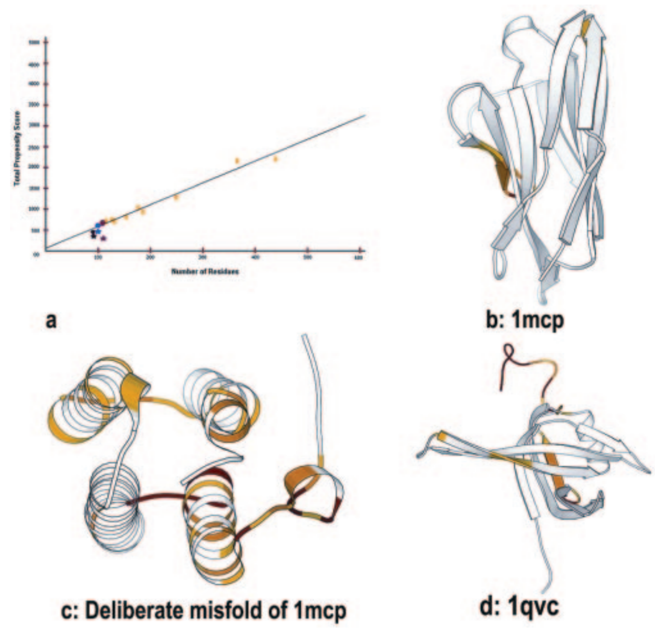

Figure 2. Structure validation using HARMONY server. (**a**) Calibration plot applied on six proteins. Actual harmony propensity score plotted as a function of protein residue length. Points marked in yellow correspond to representative members of proteins in PDB of different lengths used for calibration (Supplementary material). Circle represents correct protein models (1mcp (●), 4fd1 (●), 1y4o (●)). Asterisk represents incorrect models (Deliberate misfold of 1mcp (*), 2fd1 (*), 1tgq (*)). Incorrect models attain low scores than correct models. (**b**) HARMONY substitution scores are mapped on the immunoglobulin structure (1mcp) in relation to the reverse sequence as a control. (**c**) Same as (b) but for the deliberate misfold of 1mcp. This model has larger regions of errors in comparison to the correct model (1mcp). (**d**) Comparison of HARMONY substitution scores of actual and reverse sequence mapped on the 3D structure of *E.coli* single-stranded DNA-binding protein (1qvc). Residues 52–53, 104–110 and 115–140 are marked with different colors (red, orange and yellow) depending upon the degree of local error.

unsatisfied hydrogen bonding, buried polar/charged groups or high solvent-exposure of hydrophobic residues in protein structures (33,34). This server offers the possibility to examine local errors using amino acid substitution scores. Further, validation scores are projected on a calibration plot and mapped back on the structure for the convenient detection of gross and local errors, respectively. While, in the future, more rigorous statistical tests to assess protein structures will be included in the HARMONY approach, the present web server provides a conceptually simple means of assessment using reverse sequence. It has been shown (35) that the reverse sequence of globular proteins does not retain structural properties of the native sequence. Indeed, it has been clearly shown already (11) that use of reverse sequence serves as a simple and effective approach in the detection of errors in protein structures. HARMONY server is a simple and convenient online tool to assess the compatibility of an amino acid sequence with a proposed 3D structure and should prove useful for structural biologists, experimentalists and computational modelers.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the scientific contributions of Prof. Tom Blundell during the development of

HARMONY algorithm. N.S. and R.S. were Senior Research Fellows of the Wellcome Trust, UK. G.P. and K.S.'s stay has been supported by the Wellcome Trust. R.S. also thanks NCBS (TIFR) for infrastructural and additional financial support. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Sutcliffe,M.J., Haneef,I., Carney,D. and Blundell,T.L. (1987) Knowledge based modelling of homologous proteins, Part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng*., **1**, 377–384.
2. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*., **234**, 779–815.
3. Johnson,M.S., Srinivasan,N., Sowdhamini,R. and Blundell,T.L. (1994) Knowledge-based protein modeling. *Crit. Rev. Biochem. Mol. Biol*., **29**, 1–68.
4. Gardebien,F., Thangudu,R.R., Gontero,B. and Offmann,B. (2006) Construction of a 3D model of CP12, a protein linker. *J. Mol. Graph. Model*, (In press).
5. Zeng,J. (2000) Mini-review: computational structure-based design of inhibitors that target protein surfaces. *Comb. Chem. High. Throughput Screen*, **3**, 355–362.
6. Srinivasan,R. and Rose,G.D. (1995) LINUS: a hierarchic procedure to predict the fold of a protein. *Proteins*, **22**, 81–99.
7. Simons,K.T., Kooperberg,C., Huang,E. and Baker,D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol*., **268**, 209–225.
8. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.F. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
9. Peitsch,M.C. (2002) About the use of protein models. *Bioinformatics*, **18**, 934–938.
10. Luthy,R., Bowie,J.U. and Eisenberg,D. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.
11. Topham,C.M., Srinivasan,N., Thorpe,C.J., Overington,J.P. and Kalsheker,N.A. (1994) Comparative modelling of major house dust mite allergen Der p I: structure validation using an extended environmental amino acid propensity table. *Protein Eng*., **7**, 869–894.
12. Richardson,J.S. (2003) All-atom contacts: a new approach to structure validation. *Methods Biochem. Anal*., **44**, 305–320.
13. Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355–362.
14. Laskowski,R., MacArthur,M.W., Moss,D. and Thornton,J.M. (1993) PROCHECK—A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr*., **26**, 283–291.
15. Hooft,R.W., Vriend,G., Sander,C. and Abola,E.E. (1996) Errors in protein structures. *Nature*, **381**, 272.
16. Bujnicki,J.M., Feder,M., Rychlewski,L. and Fischer,D. (2002) Errors in the *D. radiodurans* large ribosomal subunit structure detected by protein fold-recognition and structure validation tools. *FEBS Lett*., **525**, 174–175.
17. Bujnicki,J., Rychlewski,L. and Fischer,D. (2002) Fold-recognition detects an error in the Protein Data Bank. *Bioinformatics*, **18**, 1391–1395.
18. Overington,J., Johnson,M.S., Sali,A. and Blundell,T.L. (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. Biol. Sci*., **241**, 132–145.
19. Mizuguchi,K., Deane,C.M., Blundell,T.L., Johnson,M.S. and Overington,J.P. (1998) JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623.
20. Wako,H. and Blundell,T.L. (1994) Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. *J. Mol. Biol*., **238**, 682–692.
21. Bairoch,A. and Apweiler,R. (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res*., **24**, 21–25.
22. 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.
23. Johnson,M.S., Overington,J.P. and Blundell,T.L. (1993) Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol*., **231**, 735–752.
24. Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr*., **24**, 946–950.
25. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci*., **20**, 374.
26. Thangudu,R.R., Vinayagam,A., Pugalenthi,G., Manonmani,A., Offmann,B. and Sowdhamini,R. (2005) Native and modeled disulfide bonds in proteins knowledge-based approaches toward structure prediction of disulfide-rich polypeptides. *Proteins*, **58**, 866–879.
27. Eisenberg,D., Luthy,R. and Bowie,J.U. (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol*., **277**, 396–404.
28. Colovos,C. and Yeates,T.O. (1993) Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci*., **2**, 1511–1519.
29. Willard,L., Ranjan,A., Zhang,H., Monzavi,H., Boyko,R.F., Sykes,B.D. and Wishart,D.S. (2003) VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res*., **31**, 3316–3319.
30. Melo,F., Devos,D., Depiereux,E. and Feytmans,E. (1997) ANOLEA: a www server to assess protein structures. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **5**, 187–190.
31. Matsumoto,T., Morimoto,Y., Shibata,N., Kinebuchi,T., Shimamoto,N., Tsukihara,T. and Yasuoka,N. (2000) Roles of functional loops and the C-terminal segment of a single-stranded DNA binding protein elucidated by X-Ray structure analysis. *J. Biochem. (Tokyo)*, **127**, 329–335.
32. Venclovas,C., Ginalski,K. and Kang,C. (2004) Sequence-structure mapping errors in the PDB: OB-fold domains. *Protein Sci*., **13**, 1594–1602.
33. Novotn'y,J., Rashin,A.A. and Bruccoleri,R.E. (1988) Criteria that discriminate between native proteins and incorrectly folded models. *Proteins*, **4**, 19–30.
34. Nabuurs,S.B., Spronk,C.A., Vuister,G.W. and Vriend,G. (2006) Traditional biomolecular structure determination by NMR spectroscopy allows for major errors. *PLoS Comput. Biol*., **2**, e9.
35. Karplus,K., Karchin,R., Shackelford,G. and Hughey,R. (2005) Calibrating *E*-values for hidden Markov models using reverse-sequence null models. *Bioinformatics*, **21**, 4107–4115.