WEB TOOL

Ranking the quality of protein structure models using sidechain based network properties [v1; ref status: indexed, http://f1000r.es/2eu]

Soma Ghosh[1,2], Saraswathi Vishveshwara[2]

[1]Molecular Biophysics Unit, Indian Institute of Science, Bangalore, 560012, India
[2]I.I.Sc. Mathematics Initiative, Indian Institute of Science, Bangalore, 560012, India

v1 **First published:** 21 Jan 2014, **3**:17 (doi: 10.12688/f1000research.3-17.v1)
Latest published: 21 Jan 2014, **3**:17 (doi: 10.12688/f1000research.3-17.v1)

Abstract

Determining the correct structure of a protein given its sequence still remains an arduous task with many researchers working towards this goal. Most structure prediction methodologies result in the generation of a large number of probable candidates with the final challenge being to select the best amongst these. In this work, we have used Protein Structure Networks of native and modeled proteins in combination with Support Vector Machines to estimate the quality of a protein structure model and finally to provide ranks for these models. Model ranking is performed using regression analysis and helps in model selection from a group of many similar and good quality structures. Our results show that structures with a rank greater than 16 exhibit native protein-like properties while those below 10 are non-native like. The tool is also made available as a web-server (http://vishgraph.mbu.iisc.ernet.in/GraProStr/native_non_native_ranking.html), where, 5 modelled structures can be evaluated at a given time.

Article Status Summary

Referee Responses

Referees	1	2
v1 published 21 Jan 2014	✓ report	✓ report

1 **Rahul Banerjee**, Saha Institute of Nuclear Physics India

2 **Soumen Roy**, Bose Institute India, **Rajdeep Kaur Grewal**, Bose Institute India

Latest Comments

No Comments Yet

Corresponding author: Saraswathi Vishveshwara (sv@mbu.iisc.ernet.in)

How to cite this article: Ghosh S and Vishveshwara S (2014) **Ranking the quality of protein structure models using sidechain based network properties [v1; ref status: indexed, http://f1000r.es/2eu]** *F1000Research* 2014, **3**:17 (doi: 10.12688/f1000research.3-17.v1)

Grant information: SG is supported by DBT fellowship [DBTO/BMB/SV/364], Department of Biotechnology (DBT), Government of India.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 21 Jan 2014, **3**:17 (doi: 10.12688/f1000research.3-17.v1)
First indexed: 06 May 2014, **3**:17 (doi: 10.12688/f1000research.3-17.v1)

Introduction

Proteins are known to take up unique well defined structures that allow them to function efficiently under a given condition[1]. This becomes much more fascinating when one considers the time taken by a protein to fold *in vivo*[2]. Studies over the past decades have facilitated the preparation of a blueprint of the rules that govern protein folding[3–6]. The roles of hydrophobic residues in structural packing, e.g. proline and glycine as helix breakers, are now very well established[7,8]. Details of the various pair-wise interactions that hold the structure intact are also available in the literature[9]. However, even with the wealth of resources available, determining the structure of a protein from its amino-acid sequence still remains a challenging task.

To begin with, protein structure prediction requires understanding of the differences that exist between a well-folded protein structure and a modelled structure. Many large scale decoy structures that mimic a native protein structure, but with minor variations (such as the sidechain orientations, hydrogen bonds and so on), are now freely available[10–12]. Such datasets are generated using various computational approaches such as molecular dynamics[13–15] and discrete state models[16]. Decoy structures can be compared with a large number of available native structures, hence, forming an important resource to understand patterns that are unique to natively folded proteins.

For many years now, proteins structures have been represented as networks, with residues forming nodes with edges representing various factors that are important for protein structures, such as hydrogen bonds[17], and Cα distances[18]. Although these networks help in understanding the structure of a protein at the level of secondary structures and backbone atoms, determining the subtle changes that occur at the level of sidechain interactions are not captured. We have been working on Protein Sidechain Network (PSN) for a number of years[19,20] and have done various rigorous analyses at different levels to show its usefulness[21–26]. Generating networks at the level of a sidechain not only takes care of the geometry but also the chemistry that is encoded in the sidechain atoms of every amino acid in the polypeptide chain.

Support vector machine (SVM) is a machine learning algorithm mainly used for the purpose of classification[27]. The algorithm uses a training dataset to learn patterns and finally use those patterns to classify new cases. Given the complexity of biological systems, machine learning algorithms are widely used in biology to predict cellular locations[28,29], cancer tissue classifications based on gene expression data[30–32] and further in cases of protein structures to identify SCOP classes[33], binding sites[34,35] and also the quality of protein structures using features, such as secondary structures and hydrophobicity[36,37].

Recently, we have demonstrated the capabilities of PSNs to distinguish native structures from decoy models. We started with comparing the network properties of PSNs from native and decoy models where we established the unique network features exhibited by native structures[38]. This work was further followed by an in-depth analysis, where PSNs at different interaction strengths (I_{min} = 0%–7%) and SVM were used in tandem to classify the protein as native or non-native like. Further, the method was validated using a large number of CASP 10 [10th community wide experiment on the Critical Assessment of Techniques for Protein Structure Prediction] predicted models. Overall, an accuracy of 94% was achieved by this method[39].

As an extension of our previous work, where a simple binary classification was carried out[39], here we have developed a method to rank the quality of model structures through probability estimates. This advance is particularly important in cases where one needs to select the best quality structures from a set of many similar and good quality models. Many tools have now been developed that can successfully generate many possible structure candidates from a sequence; however, predicting the best from this list is still a demanding task and needs attention. In the present study we have observed that the structures with a rank greater than 16 generally show native like properties and hence this method provides a good measure for the rank and quality of a model.

Methods

The main aim of this work was to obtain a ranking for a set of modeled structures and to select the best modeled structure that closely resembles a native structure. To achieve this goal, we obtained a large number of native and non-native structures and generated PSNs. The network parameters from the PSNs are combined with SVM to build a mathematical model and the ranking of each structure is determined using logistic regression analysis. Details of each step are provided below.

Datasets

Two sets of data were used for this study;

a) a positive dataset (PSN-QA_positive), that consisted of 5422 protein crystal structures with resolution < 3A, R-factor < 0.25 and PDB size > 100 This dataset was curated using PISCES[40],

b) a negative dataset (PSN-QA_negative) that considered different decoys as well as modelled structures from various publicly available resources and databases.

Details of the individual datasets are provided in Table 1. Finally, a total of 29543 non-native structures were obtained.

Protein Structure Network : Quality Assessment (PSN-QA)

4 Data Files

http://dx.doi.org/10.6084/m9.figshare.902838

Construction of the Protein Structure Network

As mentioned above, our laboratory has been working extensively on protein structure networks[19], specifically generated at the level of non-covalent interactions of sidechains. Details to generate PSNs are available in our previous work[20] and a brief description is provided here.

PSNs are generated by considering amino acids as nodes and edges are constructed between these nodes based on the non-covalent interaction strengths between them. Interaction strengths between any two residues is calculated as follows,

Table 1. List of resources from which decoy/modelled structures have been obtained.

Dataset	# decoy/modelled structures	Website
CASP3	971	http://predictioncenter.org/download_area/CASP3/
CASP7	10	http://predictioncenter.org/download_area/CASP7/
CASP8	10299	http://predictioncenter.org/download_area/CASP8/
CASP9	7711	http://predictioncenter.org/download_area/CASP9/
CASP10b	1428	http://predictioncenter.org/download_area/CASP10/
Rosetta protein decoy set	2660	http://depts.washington.edu/bakerpg/decoys/
Standard and complete collection of decoy set	1799	http://babylone.ulb.ac.be/decoys
Single decoy set	17	http://dd.compbio.washington.edu/download.shtml
Haemoglobin structural set	609	http://dd.compbio.washington.edu/download.shtml
Immunoglobulin structural set	3659	http://dd.compbio.washington.edu/download.shtml
Immunoglobulin structural hire set	380	http://dd.compbio.washington.edu/download.shtml

Table modified from[39]

$$I_{ij} = \frac{n_{ij} \times 100}{\sqrt{N_i \times N_j}} \qquad (1)$$

where, I_{ij} = strength of interaction between residues i and j, where $|i - j| \geq 2$; n_{ij} = number of distinct interacting atom pairs between i and j within a distance cut-off of 4.5 Å (excluding the backbone atoms); N_i and N_j are the normalization values for residues i and j obtained from a statistically significant dataset of proteins, as defined in our previous work[20]. Based on the interaction strengths between these residues, PSNs can then be generated at different interaction strength cutoffs (I_{min}), with a lower cutoff generating a dense network and including even the weaker interactions, while a higher cutoff signifies a network made of very strong non-covalent interactions and hence sparse. For this study, PSNs were generated at different I_{min}s ranging from 0% to 7%.

Various network parameters such as number of non-covalent interactions (NCov), size of the largest cluster (SLClu), clustering coefficient (CCoe), size of the largest k-1 and k-2 communities, are calculated for each PSNs generated. Furthermore, the differences between these parameters at consecutive I_{min}s are also considered in this study. In our previous studies[39], we have discussed the importance of the transition profile of the various network parameters as a function of I_{min} to characterize the native structures and therefore distinguish them from the non-native ones. Along with the network parameters, main chain hydrogen bonds (MHB)[41] were also analysed and included in the study. Table 2 provides a detailed list of all the network parameters that have been used in this study.

Support Vector Machine

As described before, SVMs are machine learning algorithms that learn patterns from a training dataset and further use that pattern to classify new datasets. In this study, we have built an SVM classifier based on the patterns that are specific for a native PSN. First, we randomly divided the datasets into a training set and a test set, so that the training set contained 3000 native structures and 3000 non-native structures. Remaining structures were set aside to form the test set. This was repeated 10 times to generate 10 random test sets and training sets. Compared to our previous study, we here went one step further and used the liblinear package of LibSVM[42,43], to obtain the probability estimates (using –s8 option in the liblinear package) of each data point and thereby to obtain ranks for each of them. Furthermore, since the different network parameter values have different ranges, the values were scaled between -10 to +10 before the analysis.

Results

Network features of PSNs

Twelve network features (at different I_{min}s) (Table 2) and MHB are combined to get a total of 94 features that best characterize a PSN. Details about these parameters and the characteristic transition curves specific to PSNs generated from native structures are discussed in detail in our previous work[39]. Briefly, the transition profiles (Figure 1) as obtained by plotting the network features of native protein structures as a function of I_{min} show three specific features, a) higher value at lower I_{min}, (b) lower value at higher I_{min}

Table 2. List of network features calculated in this study.

Parameter	Description
NCov	Number of non-covalent interactions, defined by the number of edges in a PSN
SLClu	Set of connected nodes with maximum number of residues (evaluated using DFS algorithm[44])
Top1-ComSk1	A clique is a subset of nodes in the network, such that all nodes are connected to all other nodes. Union of k-cliques such that k-1 nodes are shared between the cliques is termed as k-1-community[45]. This parameter represents the size of the largest k-1-community
Top2-ComSk1	Cumulative size of the top2 largest k-1-community
Top3-ComSk1	Cumulative size of the top3 largest k-1-community
ComSk2	Union of k-cliques such that k-2 nodes are termed as k-2-community. Represents the size of the largest k-2-community
Ccoe	Avg. clustering coefficient of the network, based on the algorithm given in[46]
CCoe-LClu	Avg. clustering coefficient of the largest cluster. This was calculated by extracting the subnetwork that forms the largest cluster
CCoe-Lcomm	Avg. clustering coefficient of the largest k-2 community
d(NCov)	Represents the transition profile of non-covalent interaction as a function of I_{min}
d(SLClu)	Represents the transition of the size of the largest cluster as a function of I_{min}
d(ComSk2)	Represents the transition of the size of the largest k-2 community as a function of I_{min}

Table adapted from[39]

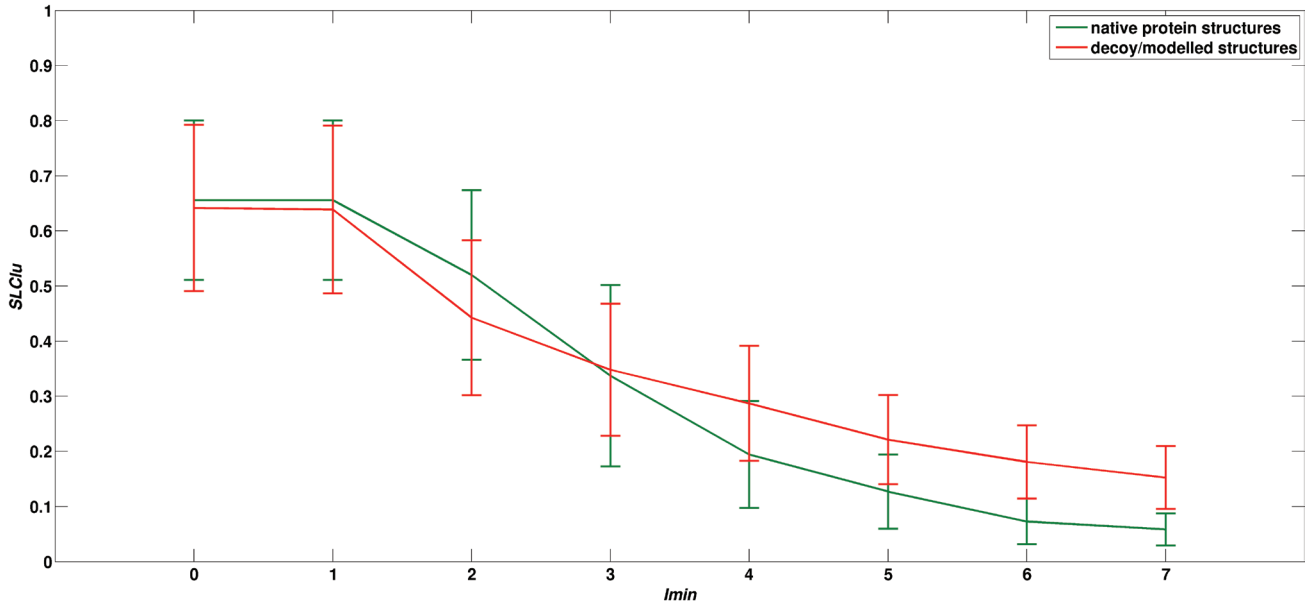


Figure 1. Transition profile of native protein structures and their corresponding decoy/modelled structures. Transition profile of one of the network features (SLClu; see Table 2) as a function of I_{min} is shown for 7 randomly selected native structures [green] and their corresponding decoy structures [red]. A clear distinction between the two transition profiles is visible, highlighting the 3 characteristic features that are uniquely displayed by native protein structures. X axis represents I_{min} from 0% to 7% and Y axis represents the average value of the SLClu obtained by native and decoy/modelled structures.

and finally (c) steep transition between I_{min} = 1%–4%. Figure 1 shows the transition profile of 7 randomly selected native protein structures and their corresponding 981 model structures. A clear difference between the transition profiles of a native protein structure and decoy/modelled structures is visible. These differences are observed in all the datasets used in this study and forms the basis of the method developed here.

SVM and the liblinear package

The main aim of this work was to obtain a ranking scheme for structure quality prediction. The 94 network features were combined into SVM using the liblinear package to obtain a ranking model. Specifically, for model generation, 'L2-regularized L2-loss ranking support vector machine' solver and cost value (c) equal to 2 was used[43]. As mentioned in the Methods section, 10 random training and test sets were obtained and the ranking model was generated for all the train sets. Finally, the model which showed the best pairwise accuracy of 98.2% was selected for further analysis.

Rank estimates

Figure 2 shows the percentage distribution of the ranks obtained by the 5422 native protein structures and 29543 non-native structures. These ranks represent the quality of the structures as determined by the network parameters using the SVM trained model. From Figure 2, it is now quite evident that native structures almost always score above 16, while the scores of the non-native structures range from -70 to 20 with the majority being ≤ 16. It should be pointed out here that the dataset of decoy structures is taken from databases such as CASP and Rosetta and therefore in many cases might also contain structures very close to native or almost native like, thereby leading to some structure scoring beyond 16, but always ≤ 20. From Figure 2, it can now be safely assumed that structures scoring above 16 show native like properties and scores of bad, unrefined models are generally very low.

Web-server

This tool is now made freely available for public use in the form of a web-server, http://vishgraph.mbu.iisc.ernet.in/GraProStr/native_non_native_ranking.html. Figure 3 shows the home page of the web-server (Figure 3a) and the output format (Figure 3b). A test case (PDB Id: 1CG5 and its decoy structures from Rosetta) is also provided with its scores as an example. Figure 4 shows the screenshot of the example test case. The tool can analyse five structures at a given time. For structures with multiple chains, individual chains are treated as different structures for the analysis. The tool accepts files in PDB formats as input and outputs the ranks for each model in a tabular format.

Discussion

Proper folding of protein structures is imparted by various energetic and topological features[1,3–9]. While the secondary structures are stabilized by backbone hydrogen bonds, the mutual orientation of the secondary structures are uniquely determined by the sidechain interactions. Although studies at the backbone level have contributed enormously to the understanding of the protein structure[17,18], they are not sufficient to understand the subtle balance

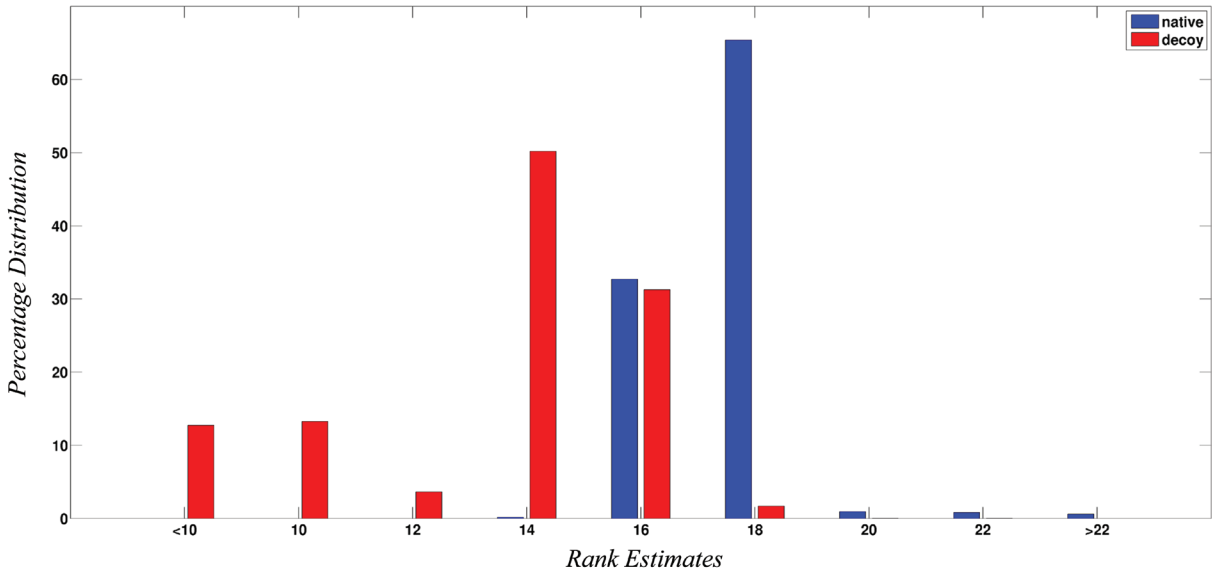
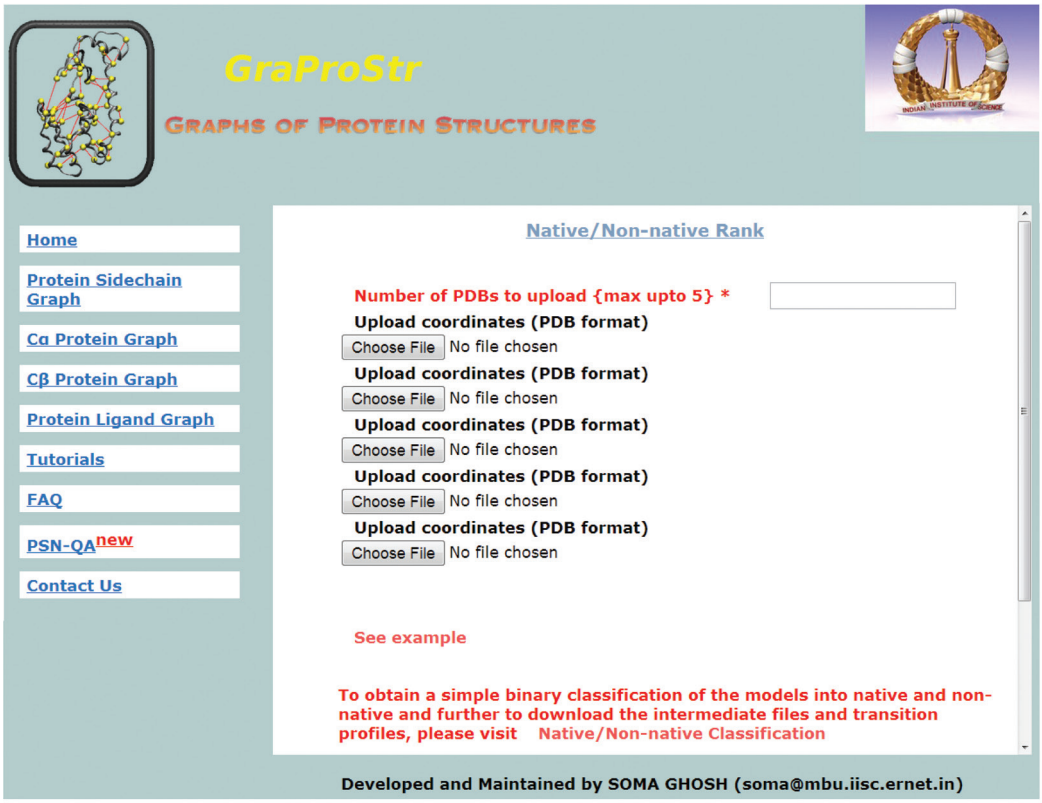


Figure 2. Percentage distribution of rank for native and non-native structures. The figure shows the percentage distribution of ranks for the 5422 native structures (blue) and 29543 decoy/modelled structures (red). X-axis represents ranks while Y-axis represents the percentage distribution. It is clear that native structures have higher ranks (> 16) as compared to the decoy/modelled structures.

a)



b)

GraProtS

3 pdb files uploaded for analysis :

1FKB_l10020.pdb : renamed as pdb1.pdb

1AIU.pdb : renamed as pdb2.pdb

1ASH.pdb : renamed as pdb3.pdb

Checking PDB file for atom coordinates

Number of chains (pdb number 1): No chain information provided .. Assuming monomeric structure

Number of chains (pdb number 2): 1

Number of chains (pdb number 3): 1

PDB pruning...

Running.....

Structural Assessment of the PDB structures uploaded.

PDB_id(with_chain)	Rank
pdb2A.pdb	18.4595
pdb3A.pdb	17.5152
pdb1.pdb	5.4147

Figure 3. Web-server for ranking protein structures. The figure shows screenshots of the **a**) home page and **b**) results page for structure ranking. At a given time, 5 structures can be uploaded. For structures with multiple chains, each chain would be treated individually. The output would be provided in a tabular format.

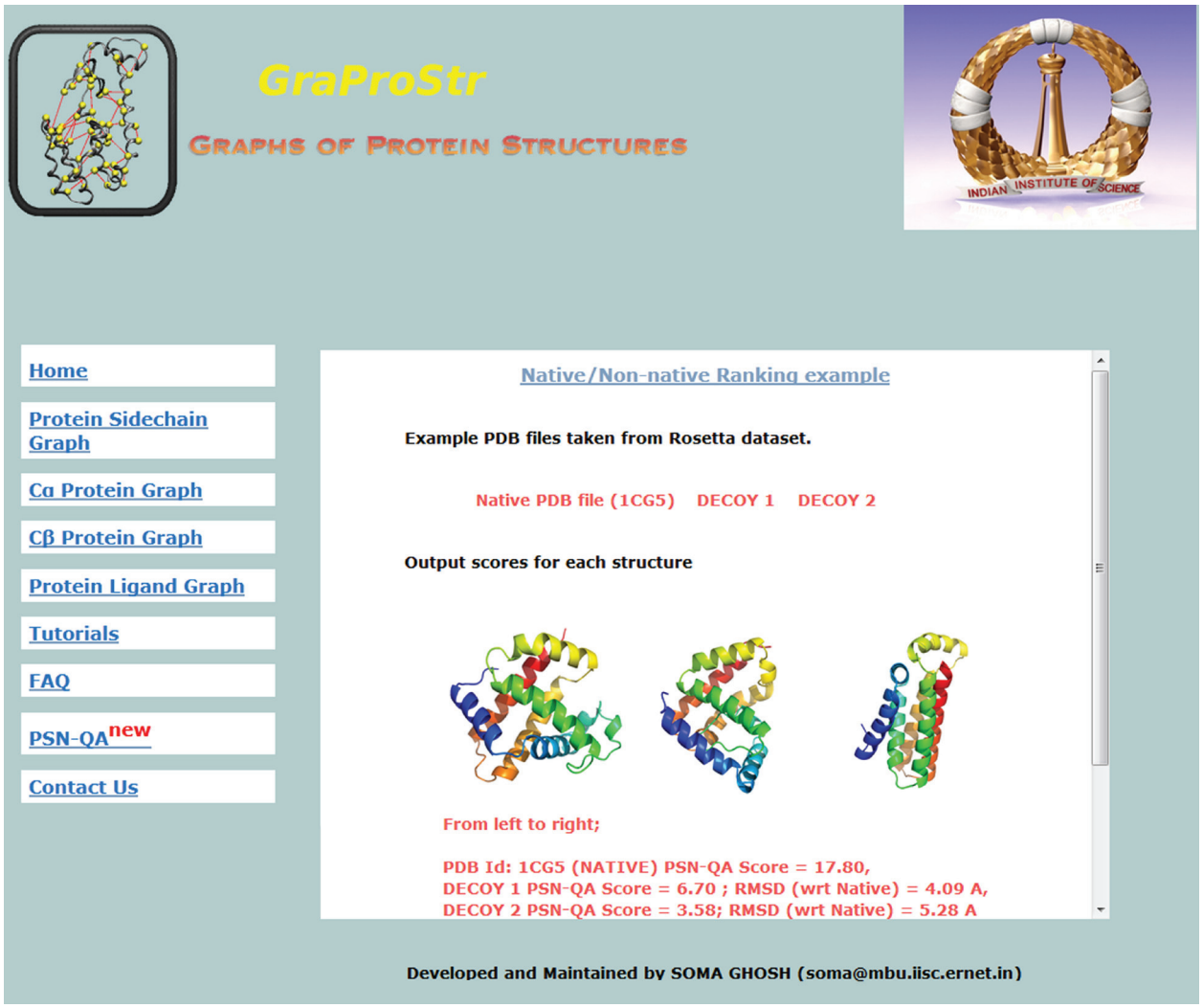


Figure 4. Example test case as shown in the web-server. For easy understanding, a test case of native structure (PDB Id: 1CG5) and its two decoy structures (from Rosetta) is also made available. The page shows the structures and the PSN scores obtained by them. PDB files are also available for download.

at the atomic level. Our previous studies have highlighted the role of non-covalent interactions of the sidechain atoms in functioning[23,25,26] as well as stability[22,24] of protein structures. Protein structure networks are designed to account for sidechain interactions and therefore the network captures not only the geometric but also the chemistry encoded in the sidechain.

In our earlier studies, we had exploited protein structure networks to discriminate the native structures from the non-native ones. This is mainly done at the level of sidechain with only one important feature, MHB, representing the properties of the backbone atoms. In all these studies[38,39], discrimination between the two sets is done qualitatively, with the method simply classifying the structures as native or non-native. Such qualitative analysis becomes ineffective when used for closely related and almost native like structures. However, given the current state of art in the field of protein structure prediction, we believe that expertise has been attained to predict near native like structures and more work is required now to select the best structure from a set of very similar structures.

The present work is an extension of our earlier work, where we have addressed the issue described above in a quantitative manner. Here, we have built a model that would score the structures based on how closely they mimic a native structure, instead of providing a simple binary classification. We were able to use the liblinear package of libSVM to build such a model. The model was further tested on a set of 5422 native structures and 29543 decoy/modelled structures. The ranking scheme (Figure 2) is clearly able to discriminate good structures from the bad ones. All the 5422 native

structures get a rank greater than 16, while the scores for decoy/modelled structures range from -70 to 20. Overall, it can be concluded that structures with score > 16 display native like properties as evaluated from a network perspective and the models below the score of 12 are definitely show non-native like properties and do not mimic native structures.

Conclusion

In summary, large numbers of native as well as decoy/modelled structures have been used to build an SVM model. This model was trained using 94 features that included 93 network parameters and main chain hydrogen bonds. The model has an overall accuracy of 98.2% and can successfully rank structures based on their quality as determined from protein structure networks. Generally, structures with rank > 16 display native like properties and can be regarded as good quality structures. This is an important advancement from the previous qualitative assessments and would be helpful in cases where one needs to extract the best structure from a set of closely related structures.

Data and software availability

Data

Figshare: Protein Structure Network : Quality Assessment (PSN-QA), doi: 10.6084/m9.figshare.902838[47].

Software

Protein Structure Network Quality Assessment (PSN-QA) tool: http://vishgraph.mbu.iisc.ernet.in/GraProStr/native_non_native_ranking.html

Author contributions

SV conceptualized the idea and supervised the project. SG performed the analysis and developed the web server. Both authors wrote and approved the final manuscript.

Competing interests

No competing interests were disclosed.

Grant information

SG is supported by DBT fellowship [DBTO/BMB/SV/364], Department of Biotechnology (DBT), Government of India.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We thank Subhojyoti Chatterjee who participated in the earlier work of network analysis on decoy structures.

References

1. Anfinsen C: **Principles that govern the folding of protein chains.** *Science.* 1973; **181**(4096): 223–230.
PubMed Abstract | Publisher Full Text
2. Daggett V, Fersht A: **The present view of the mechanism of protein folding.** *Nat Rev Mol Cell Biol.* 2003; **4**(6): 497–502.
PubMed Abstract | Publisher Full Text
3. Chan HS, Dill KA: **The protein folding problem.** *Phys Today.* 1993; **46**(2): 24.
Publisher Full Text
4. Fersht AR: **From the first protein structures to our current knowledge of protein folding: delights and scepticisms.** *Nat Rev Mol Cell Biol.* 2008; **9**(8): 650–654.
PubMed Abstract | Publisher Full Text
5. Karplus M: **Behind the folding funnel diagram.** *Nat Chem Biol.* 2011; **7**(7): 401–404.
PubMed Abstract | Publisher Full Text
6. Rose GD, Fleming PJ, Banavar JR, *et al.*: **A backbone-based theory of protein folding.** *Proc Natl Acad Sci U S A.* 2006; **103**(45): 16623–16633.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Dyson HJ, Wright PE, Scheraga HA: **The role of hydrophobic interactions in initiation and propagation of protein folding.** *Proc Natl Acad Sci U S A.* 2006; **103**(35): 13057–13061.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Kellis JT Jr, Nyberg K, Sali D, *et al.*: **Contribution of hydrophobic interactions to protein stability.** *Nature.* 1988; **333**(6175): 784–786.
PubMed Abstract | Publisher Full Text
9. Burley S, Petsko G: **Aromatic-aromatic interaction: a mechanism of protein structure stabilization.** *Science.* 1985; **229**(4708): 23–28.
PubMed Abstract | Publisher Full Text
10. Gilis D: **Protein decoy sets for evaluating energy functions.** *J Biomol Struct Dyn.* 2004; **21**(6): 725–736.
PubMed Abstract | Publisher Full Text
11. Samudrala R, Levitt M: **Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction.** *Protein Sci.* 2000; **9**(7): 1399–1401.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Tsai J, Bonneau R, Morozov AV, *et al.*: **An improved protein decoy set for testing energy functions for protein structure prediction.** *Proteins.* 2003; **53**(1): 76–87.
PubMed Abstract | Publisher Full Text
13. Huang ES, Subbiah S, Tsai J, *et al.*: **Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations.** *J Mol Biol.* 1996; **257**(3): 716–725.
PubMed Abstract | Publisher Full Text
14. Zhang J, Zhang Y: **A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction.** *PLoS One.* 2010; **5**(10): e15386.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Wu S, Skolnick J, Zhang Y: **Ab initio modeling of small proteins by iterative TASSER simulations.** *BMC Biol.* 2007; **5**: 17.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Park B, Levitt M: **Energy functions that discriminate X-ray and near-native folds from well-constructed decoys.** *J Mol Biol.* 1996; **258**(2): 367–392.
PubMed Abstract | Publisher Full Text
17. Rahat O, Alon U, Levy Y, *et al.*: **Understanding hydrogen-bond patterns in proteins using network motifs.** *Bioinformatics.* 2009; **25**(22): 2921–2928.
PubMed Abstract | Publisher Full Text
18. Atilgan AR, Akan P, Baysal C: **Small-world communication of residues and significance for protein dynamics.** *Biophys J.* 2004; **86**(1 Pt 1): 85–91.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Vishveshwara S, Brinda K, Kannan N: **Protein structure: insights from graph theory.** *J Theor Comput Chem.* 2002; **1**(01): 187–211.
Publisher Full Text
20. Kannan N, Vishveshwara S: **Identification of side-chain clusters in protein structures by a graph spectral method.** *J Mol Biol.* 1999; **292**(2): 441–464.
PubMed Abstract | Publisher Full Text
21. Sathyapriya R, Vijayabaskar M, Vishveshwara S: **Insights into Protein–DNA Interactions through structure network analysis.** *PLoS Comput Biol.* 2008; **4**(9): e1000170.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Deb D, Vishveshwara S, Vishveshwara S: **Understanding protein structure from a percolation perspective.** *Biophys J.* 2009; **97**(6): 1787–1794.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Bhattacharyya M, Vishveshwara S: **Probing the allosteric mechanism in pyrrolysyl-tRNA synthetase using energy-weighted network formalism.** *Biochemistry.* 2011; **50**(28): 6225–6236.
PubMed Abstract | Publisher Full Text
24. Brinda K, Vishveshwara S: **A network representation of protein structures: implications for protein stability.** *Biophys J.* 2005; **89**(6): 4159–4170.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Brinda KV, Kannan N, Vishveshwara S: **Analysis of homodimeric protein interfaces by graph-spectral methods.** *Protein Eng.* 2002; **15**(4): 265–277.
PubMed Abstract | Publisher Full Text

26. Kannan N, Vishveshwara S: **Aromatic clusters: a determinant of thermal stability of thermophilic proteins.** *Protein Eng.* 2000; **13**(11): 753–761.
PubMed Abstract | Publisher Full Text

27. Noble WS: **What is a support vector machine?** *Nat Biotechnol.* 2006; **24**(12): 1565–1567.
PubMed Abstract | Publisher Full Text

28. Park KJ, Kanehisa M: **Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs.** *Bioinformatics.* 2003; **19**(13): 1656–1663.
PubMed Abstract | Publisher Full Text

29. Hua S, Sun Z: **Support vector machine approach for protein subcellular localization prediction.** *Bioinformatics.* 2001; **17**(8): 721–728.
PubMed Abstract | Publisher Full Text

30. Khan J, Wei JS, Ringner M, *et al.*: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.** *Nat Med.* 2001; **7**(6): 673–679.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Guyon I, Weston J, Barnhill S, *et al.*: **Gene selection for cancer classification using support vector machines.** *Machine Learn.* 2002; **46**(1–3): 389–422.
Publisher Full Text

32. Furey TS, Cristianini N, Duffy N, *et al.*: **Support vector machine classification and validation of cancer tissue samples using microarray expression data.** *Bioinformatics.* 2000; **16**(10): 906–914.
PubMed Abstract | Publisher Full Text

33. Cai YD, Liu XJ, Xu Xb, *et al.*: **Prediction of protein structural classes by support vector machines.** *Comput Chem.* 2002; **26**(3): 293–296.
PubMed Abstract

34. Koike A, Takagi T: **Prediction of protein–protein interaction sites using support vector machines.** *Protein Eng Des Sel.* 2004; **17**(2): 165–173.
PubMed Abstract | Publisher Full Text

35. Bradford JR, Westhead DR: **Improved prediction of protein–protein binding sites using a support vector machines approach.** *Bioinformatics.* 2005; **21**(8): 1487–1494.
PubMed Abstract | Publisher Full Text

36. Kryshtafovych A, Fidelis K: **Protein structure prediction and model quality assessment.** *Drug Discovery Today.* 2009; **14**(7): 386–393.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Dong Q, Chen Y, Zhou S: **A machine learning-based method for protein global model quality assessment.** *Int J Gen Syst.* 2011; **40**(04): 417–425.
Publisher Full Text

38. Chatterjee S, Bhattacharyya M, Vishveshwara S: **Network properties of protein-decoy structures.** *J Biomol Struct Dyn.* 2012; **29**(6): 606–622.
PubMed Abstract | Publisher Full Text

39. Chatterjee S, Ghosh S, Vishveshwara S: **Network properties of decoys and CASP predicted models: A comparison with native protein structures.** *Mol Biosyst.* 2013; **9**(7): 1774–1788.
PubMed Abstract | Publisher Full Text

40. Wang G, Dunbrack RL: **PISCES: a protein sequence culling server.** *Bioinformatics.* 2003; **19**(12): 1589–1591.
PubMed Abstract | Publisher Full Text

41. McDonald I, Naylor D, Jones D, *et al.*: **HBPLUS computer program**. *Department of Biochemistry and Molecular Biology, University College, London, UK.* 1993.
Reference Source

42. Chang CC, Lin CJ: **LIBSVM: a library for support vector machines.** *ACM Trans Intell Syst Technol (TIST).* 2011; **2**(3): 27.
Publisher Full Text

43. Fan RE, Chang KW, Hsieh CJ, *et al.*: **LIBLINEAR: A library for large linear classification.** *J Machine Learn Res.* 2008; **9**: 1871–1874.
Publisher Full Text

44. Leiserson CE, Rivest RL, Stein C, *et al.*: **Introduction to algorithms**. The MIT press. 2001.
Reference Source

45. Adamcsek B, Palla G, Farkas IJ, *et al.*: **CFinder: locating cliques and overlapping modules in biological networks.** *Bioinformatics.* 2006; **22**(8): 1021–1023.
PubMed Abstract | Publisher Full Text

46. Soffer SN, Vázquez A: **Network clustering coefficient without degree-correlation biases.** *Phys Rev E Stat Nonlin Soft Matter Phys.* 2005; **71**(5 Pt 2): 057101.
PubMed Abstract | Publisher Full Text

47. Ghosh S, Vishveshwara S: **Protein Structure Network: Quality Assessment (PSN-QA).** *Figshare.* 2014.
Data Source

Current Referee Status:

Referee Responses for Version 1

Soumen Roy, **Rajdeep Kaur Grewal**
Bose Institute, Kolkata, India

Approved: 06 May 2014

Referee Report: 06 May 2014
This paper aims at ranking protein structures in order to differentiate native protein structures from non-native/decoy models. For this, the authors employ machine learning approaches (Support Vector Machines) and assign ranks, based on regression analysis, to these models using Protein Structure Networks. This study includes the side chain interactions of amino acid residues unlike the previous network based approaches for detecting the most native-like structures from a huge set of decoys.

The authors have built upon their earlier work employing Protein Structure Networks (PSN) to differentiate the native conformations from decoy/modelled structures; PSN parameters (93 network features) along with main chain hydrogen bonds were used to built the SVM classifier. The web tool provides wide and simple accessibility of the aforementioned methods to the larger community. Non-specialists should find it useful.

We have the following comments:

- Use of network metrics – mostly based on “size” of higher communities and largest cluster, and the average clustering coefficients have been discussed by the authors. A discussion as to why these particular metrics have been chosen over so many other available network metrics would certainly be helpful for researchers with a keen interest in network theory.

- The methodology of ranking structures could certainly be presented in more detail. The part preceding *“Finally, the model which showed the best pairwise accuracy of 98.2% was selected for further analysis”* would do much better with a more detailed explanation, especially about how LibSVM is useful here.

It would be good to know the computational complexity as well as advantages of the present approach over other accessible packages. A table summarizing this would be a highly desirable addition especially because the authors state in the Conclusion that this is an important advancement from the previous qualitative assessments and would be helpful in cases where one needs to extract the best structure from a set of closely related structures.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Rahul Banerjee
Crystallography and Molecular Biology Division, Saha Institute of Nuclear Physics, Kolkata, India

Approved: 14 April 2014

Referee Report: 14 April 2014

- The title is appropriate, and the abstract represents a suitable summary of the work.
- Although the design and methodology of the calculation are appropriate for the subject under study, some aspect of the calculation could have been explained in greater detail. This is discussed in greater detail in the report below.
- The conclusions are justified on the basis of the results obtained in the study.
- Enough information has been provided to replicate the calculations.

The authors provide a novel method to validate protein structures based on the network properties of non-bonded side chain contacts within proteins. The method could find extensive application in the structural validation of both experimentally determined protein structures by x-ray crystallography, or modeled structures. Thus as a validation tool it could prove to be an extremely valuable addition to other existing methods. The authors have also installed a web server, thus making the facility available to a wide cross section of potential users.

The success of any validation method depends on the scoring (or ranking) scheme adopted to sort structures based on some criteria. Unfortunately, the details regarding the ranking scheme are extremely terse or assume that the reader will be conversant with the details of support vector machines (SVM) and the relevant software (LibSVM). That need not be the case, as potentially work such as this should have a wide appeal. Although the authors do cite previous work, they could discuss this in somewhat greater detail. What do the terms or options used in the sentence '*Specifically, for model generation, 'L2-regularized L2-loss ranking support vector machine' solver and cost value (c) equal to 2 was used*.', actually signify? Why was the specific option (s8) chosen?

The authors could also compare their methodology with currently available validation packages such as Procheck or Molprobity on a small database consisting of native protein and decoys. Given the fact that experimentally determined erroneous structures occasionally seep through the currently available validation filters, this method could provide crucial information in error detection, where other methods consistently fail.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.