WEB TOOL

Ranking the quality of protein structure models using sidechain based network properties [version 1; referees: 2 approved]

Soma Ghosh\textsuperscript{1,2}, Saraswathi Vishveshwara\textsuperscript{2}

\textsuperscript{1}Molecular Biophysics Unit, Indian Institute of Science, Bangalore, 560012, India
\textsuperscript{2}I.I.Sc. Mathematics Initiative, Indian Institute of Science, Bangalore, 560012, India

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.

This article is included in the Machine learning: life sciences collection.

Corresponding author: Saraswathi Vishveshwara (sv@mbu.iisc.ernet.in)
Competing interests: No competing interests were disclosed.

How to cite this article: Ghosh S and Vishveshwara S. Ranking the quality of protein structure models using sidechain based network properties [version 1; referees: 2 approved] F1000Research 2014, 3:17 (doi: 10.12688/f1000research.3-17.v1)

Copyright: © 2014 Ghosh S and Vishveshwara S. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

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.

Introduction

Proteins are known to take up unique well defined structures that allow them to function efficiently under a given condition. This becomes much more fascinating when one considers the time taken by a protein to fold in vivo. Studies over the past decades have facilitated the preparation of a blueprint of the rules that govern protein folding. The roles of hydrophobic residues in structural packing, e.g. proline and glycine as helix breakers, are now very well established. Details of the various pair-wise interactions that hold the structure intact are also available in the literature. 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. Such datasets are generated using various computational approaches such as molecular dynamics and discrete state models. 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, and Cα distances. 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 and have done various rigorous analyses at different levels to show its usefulness. 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. 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, cancer tissue classifications based on gene expression data, and further in cases of protein structures to identify SCOP classes, binding sites and also the quality of protein structures using features, such as secondary structures and hydrophobicity.

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. This work was further followed by an in-depth analysis, where PSNs at different interaction strengths (Imin = 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 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.

As an extension of our previous work, where a simple binary classification was carried out, 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 any 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 modelled 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 < 3Å, R-factor < 0.25 and PDB size > 100 This dataset was curated using PISCES.

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, specifically generated at the level of non-covalent interactions of sidechains. Details to generate PSNs are available in our previous work 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.

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

Table modified from

\[
I_{ij} = \frac{n_{ij} \times 100}{\sqrt{N_i \times N_j}}
\]

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. 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, 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) 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, 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. 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.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCov</td>
<td>Number of non-covalent interactions, defined by the number of edges in a PSN</td>
</tr>
<tr>
<td>SLClu</td>
<td>Set of connected nodes with maximum number of residues (evaluated using DFS algorithm)</td>
</tr>
<tr>
<td>Top1-ComSk1</td>
<td>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. This parameter represents the size of the largest k-1-community</td>
</tr>
<tr>
<td>Top2-ComSk1</td>
<td>Cumulative size of the top2 largest k-1-community</td>
</tr>
<tr>
<td>Top3-ComSk1</td>
<td>Cumulative size of the top3 largest k-1-community</td>
</tr>
<tr>
<td>ComSk2</td>
<td>Union of k-cliques such that k-2 nodes are termed as k-2-community. Represents the size of the largest k-2-community</td>
</tr>
<tr>
<td>Ccoe</td>
<td>Avg. clustering coefficient of the network, based on the algorithm given in</td>
</tr>
<tr>
<td>Ccoe-LClu</td>
<td>Avg. clustering coefficient of the largest cluster. This was calculated by extracting the subnetwork that forms the largest cluster</td>
</tr>
<tr>
<td>Ccoe-Lcomm</td>
<td>Avg. clustering coefficient of the largest k-2 community</td>
</tr>
<tr>
<td>d(NCov)</td>
<td>Represents the transition profile of non-covalent interaction as a function of I_{min}</td>
</tr>
<tr>
<td>d(SLClu)</td>
<td>Represents the transition of the size of the largest cluster as a function of I_{min}</td>
</tr>
<tr>
<td>d(ComSk2)</td>
<td>Represents the transition of the size of the largest k-2 community as a function of I_{min}</td>
</tr>
</tbody>
</table>

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 $I_{\text{min}} = 1\% - 4\%$. 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$^1$. 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](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 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.
at the atomic level. Our previous studies have highlighted the role of non-covalent interactions of the sidechain atoms in functioning\textsuperscript{23,25,26} as well as stability\textsuperscript{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\textsuperscript{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

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


Open Peer Review

Current Referee Status: ✔ ✔

Version 1

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

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 build 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.

Competing Interests: No competing interests were disclosed.

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.
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.

**Competing Interests:** No competing interests were disclosed.

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.
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com