DISCRIMINATION OF NATIVE FOLDS USING NETWORK PROPERTIES OF PROTEIN STRUCTURES

ALPER KÜÇÜKURAL
Faculty of Engineering and Natural Sciences,
Sabanci University, Orhanli, Tuzla, Istanbul, Turkey

O. UĞUR SEZERMAN
Faculty of Engineering and Natural Sciences,
Sabanci University, Orhanli, Tuzla, Istanbul, Turkey

AYTÜL ERÇİL
Faculty of Engineering and Natural Sciences,
Sabanci University, Orhanli, Tuzla, Istanbul, Turkey

Graph theoretic properties of proteins can be used to perceive the differences between correctly folded proteins and well designed decoy sets. 3D protein structures of proteins are represented with graphs. We used two different graph representations: Delaunay tessellations of proteins and contact map graphs. Graph theoretic properties for both graph types showed high classification accuracy for protein discrimination. Fisher, linear, quadratic, neural network, and support vector classifiers were used for the classification of the protein structures. The best classifier accuracy was over 98%. Results showed that characteristic features of graph theoretic properties can be used in the detection of native folds.

1 Introduction

Proteins are the major players responsible for almost all the functions within the cell. Protein function, moreover, is mainly determined by its structure. Several experimental methods already exist to obtain the protein structure, such as x-ray crystallography and NMR. All of these methods, however, have their limitations: they are neither cost nor labor effective. Therefore, an imminent need arises for computational methods that determine protein structure which will reveal clues about the mechanism of its function. Determining the rules governing protein function will enable us to design proteins for specific function and types of interactions. [1] This course of action has vast application areas ranging from the environmental to the pharmaceutical industries. Additionally, these designed proteins should have native like protein properties to perform their function without destabilizing under physiological conditions.

There are several methods developed to find the three dimensional structure of proteins. Since these models are created by computer programs their overall structural properties may differ from those of native proteins. There is a need for distinguishing near native like structures (accurate models) from those that do not show native like structural properties. This paper aims to define a function that can distinguish the native protein

structures from artificially generated non native like protein structures. The proposed function can also be used in the protein folding problem as well as domain recognition and structural alignment of proteins.

2 Methods

The evaluation function consists of two parts: the network properties of the graphs obtained from the proteins and the contact potentials. Graphs are employed to solve many problems in protein structure analysis as a representation method. [2, 3] Protein structure can be converted into a graph where the nodes represent the C_α atoms of the residues and the links between them represent interactions (or contacts) between these residues.

The two most commonly used representations of 3D structures of proteins in graph theory are contact maps and Delaunay tessellated graphs [4, 5]. Both graphs can be represented as an N×N matrix S for a protein which has N residues. If the residues are in contact the $S_{i,j}=1$, otherwise $S_{i,j}=0$ [6, 7]. Contact definition differs for both graphs. In contact map, if the distance between C_α atoms of residues i and j is smaller than a cut-off value then they are considered to be in contact. Several distances ranging from 6.5 Aº to 8 Aº have been used in the literature. 6.8 Aº has been found to be a good definition of a contact between residues, therefore in our work we used 6.8 Aº as the contact cut off value [5].

On the other hand, Delaunay tessellated graphs consist of partitions produced between a set of points. A point is represented by an atom position in the protein for each residue. This atom position can be chosen as α carbon, β carbon or the center of mass of the side chain. There is a certain way to connect these points by edges so as to have Delaunay simplices which form non-overlapping tetrahedrals [4]. A Delaunay tessellated graph includes the neighborhood (contact) information of these Delaunay simplices. In this work, we used Qhull program to derive the Delaunay tessellated graph of our proteins using the alpha carbon atoms as simplices [8, 21].

Several network properties of the graphs are employed to distinguish the graphs of native proteins from those obtained from artificially created near native conformations, called decoy sets. The first network property is the degree or connectivity k which is the number of edges incident of a vertex i [4]. The average degree of a protein structure is calculated by the mean of the degree distribution of the graph. If the average degree is high, this points out to a globular structure where many residues establish many contacts with each other. Unfolded proteins would have very low average degree value. Natural proteins folds are compact, and measures using the compactness of the proteins can distinguish the native folds from those of artificially generated decoy set. The second graph property is the second connectivity which is calculated by the sum of the contacts of each neighbor of a node. The second connectivity is a measure we defined that also shows the compactness of the graph. If the structure is composed of small compact domains rather than one globular structure, the structure would have high average degree

but low second connectivity numbers. The attractiveness of this value is its ability to distinguish such structures.

The third graph property is the clustering coefficient which measures how well the neighbors are connected to each other, thus forming a network of contacts (clique). The clustering coefficient C for each node is calculated by

$$C_n = \frac{2E_n}{k(k-1)}$$

where En is the actual edges between the neighbors of the residue n and k is the degree. If all the neighbors of a node i are connected to each other, then they form a tight clique and the C_i value becomes 1. The clustering coefficient of the graph C is the average of all the C_n values [4, 9].

Graph properties can only capture overall structural properties of the proteins but do not measure physiochemical interactions between the atoms that are in contact in the folded form. The second part of the evaluation function uses contact potentials to capture the favorability of physicochemical interactions between the contacting residues of the folded protein. Contact potentials are statistical potentials that are calculated from experimentally known 3D structures of proteins which calculate the frequencies of occurrences of all possible contacts and convert them into energy values so that frequently occurring contacts have favorable contact scores. This method is an approximation to actual physico-chemical potentials but they have been shown to work as target energy functions on the protein folding problem [7, 8, 12, 13].

In this study, the average contact potential scores were calculated using contact potential matrix by Jernigan *et. al.* [10]. There are other contact potential matrices that are widely used as well [11], since they are highly correlated with each other, we found it sufficient to use Jernigan matrix to see the discriminative power of contact potentials in our problem. The degree, clustering coefficient, second connectivity and their moments along with Jernigan potential scores are employed as dimensions of the classification methods. Using the average values causes loss of information on the distribution of each variable; therefore we used moments to better capture the distributions of all the features.

Several classification methods are used to find out whether the graph theoretic properties can discriminate the native proteins while determining which graph representation and data classification method yields the best results.

3 Background and Related Works

Several attempts have been made to define a function to distinguish native folds from incorrectly folded proteins. In early studies, Novotny *et. al.* looked at various concepts such as solvent-exposed side-chain non-polar surface, number of buried ionizable groups, and empirical free energy functions that incorporate solvent effects for ability to discriminate between native folds and those misfolded ones in 1988 [25]. Vajda *et. al.*

used combination of hydrophobic folding energy and the internal energy of proteins which showed importance of relaxation of bond lengths and angles contributing to the internal energy terms in detection of native folds [2, 22].

McConkey *et. al.* have used contact potentials as well to distinguish native proteins. They calculated the contacts from Voronoi tessellated graphs of the native proteins and the decoy sets. They assumed a normal distribution of contact energy values and calculated the z scores to show if the native protein has a very high z-score compared to z-score of the decoy structures (or the contact energy of the native structure ranks high compared to decoy structures created for that structure). The scoring function can effectively distinguish 90% of the native structures on several decoy sets created from native protein structures [14].

Another scoring function derived by Wang *et. al.* is based on calculating distances (RMSD) between all the Cα atoms in native proteins and other conformations in given decoy sets. They show their function distinguish better than other functions depending on the quality of the decoy sets [15].

Beside the knowledge based potentials, approximate free energy potentials are also used to discriminate native proteins by Gatchel *et. al.* [15]. In their approach they defined a free energy potential that combines molecular mechanics potentials with empirical solvation and entropic terms. Their free energy potential's discrimination power improved when the internal energy of the structure was added to the solvation energy. [16]

The hydrophobic effect on protein folding and its importance to discrimination of proteins is also stated by Fain *et. al.* Their approach is based on discovering optimal hydrophobic potentials for this specific problem, by using different optimization methods. [17]

Using graph properties to distinguish native folds was first done by Taylor *et. al.* They state that using degree, clustering coefficient, and the average path length information can help distinguish native proteins. They determine a short list based on these properties. The natives' appearance in the short list indicates that these properties can distinguish the native like structures. Of 43 structures set in which they worked, the native was placed in the short list in 27 of them. [4]

All of the previous works do not treat the problem as a classification problem; they only check whether the native structure ranks high according to their scoring scheme. Several classification and clustering methods such as neural network based approaches and support vector machines have been widely used in other successful applications related to protein structure. The success of the classification depends on the features that are used to discriminate the classes [7, 18, 19].

In this paper we use combination of contact potentials (to capture the physicochemical interactions between the contacting residues that are formed upon folding) and network properties of the graph (which shows compactness of the structure). Using these values as the feature vectors, we used several classification methods to distinguish native and decoy protein classes.

4 Dataset

The first data set employed in the experiments, which is from PISCES database[20], has 1364 non-homologous proteins, and their resolution < 2.2Å, crystallographic R factor < 0.23, and maximum pair wise sequence identity < 30%. The second data set consists of 1364 artificially generated and well designed decoy set; the third one is 101 artificially generated straight helices. Decoy sets are generated by randomly locating C_α atoms at about 3.83A° distance while avoiding the self-intersection of C_α atoms and keeping the globular structure approximately at the same size and shape of an average protein [4]. Further details of decoy set generation stage can be found in the article of Wang *et. al.* [26].

The feature values in the data set possessed large variations in some cases. Therefore, to see the impact of outliers in classification accuracy, we performed a simple outlier analysis technique based on the elimination of all the values that are three standard deviations away from the mean for the given data set. Approximately 9% of the data was eliminated for each dataset.

5 Results

Average degree, clustering coefficient, second connectivity are used as structural features. Besides the averages for the properties, moments of the probability distributions were calculated for each property such as standard deviation, skewness and kurtosis of the distributions whereas skewness measures the asymmetry of the distribution and kurtosis measures the "peakedness" of the distribution. Average Jernigan potential scores are given as sequence dependent energy features. These features are supplied as input vector to several classification methods in PRTools [19]. We first tested which graph representation method is more suitable for the given problem. The results from Delaunay tessellated graphs and contact map results are given in Table 1. The contact map had much better prediction accuracy since it captures actual compactness information of the protein structure. In some cases, tessellated graphs may represent the distant residues as if they are in close contact; this representation may be the reason for the difference in classification accuracy.

We randomly selected half of the data five times and performed a five fold cross validation on each data set to reduce to run time for the classifiers especially for the support vector classifier. The classification accuracy and two standard deviation neighborhood of these values are shown in the tables.

Table 1. indicates that the best classification accuracy was obtained from normal density based quadratic classifier (qdc) [19]. Even though some of the other classifiers performed very close to the qdc, we proceeded to focus on qdc for the rest of the paper. Table 1. also shows that outlier analysis improved the results by a minimum of 1 % independent of the classification method used.

We optimized the SVM results using kernel parameters (σ) and regularization parameters (C) for each of the kernel function separately. Changing the regularization parameter (C) did not affect classification error rates. After parameter optimization the best results from SVM were obtained when the polynomial kernel was used with while σ was 2.

Table 1. Classification accuracy table using all the features including the moment values

	Contact Maps		Delaunay Tes.	
Classifier	After OA	Before OA	After OA	Before OA
Support vector class.	98.02%± 0.44	96.47%± 0.93	94.78%± 1.62	93.56%± 1.12
Norm. dens. based linear	98.72%± 0.53	97.12%± 1.02	94.85%± 1.67	93.41%± 0.94
Norm. dens. based quad.	98.87%± 0.49	98.08%± 1.32	94.81%± 1.20	92.91%± 0.52
Binary decision tree	95.61%± 1.97	94.04%± 1.88	85.77%± 2.01	82.23%± 4.17
Quadratic classifier	98.54%± 0.71	98.11%± 0.88	94.97%± 1.13	93.51%± 0.74
Linear perceptron	95.28%± 1.56	93.98%± 1.13	50.46%±10.81	54.46%± 8.53
Random neural network	96.76%± 0.76	95.40%± 1.72	88.81%± 2.27	86.10%± 2.13
k-nearest neighbor (k=3)	97.67%± 1.26	95.93%± 0.98	85.06%± 0.82	83.95%± 2.32
Parzen classifier	97.04%± 0.86	95.25%± 1.12	85.89%± 2.43	84.51%± 2.94
Parzen density based	98.59%± 0.56	97.12%± 1.77	88.62%± 3.08	86.66%± 2.71
Naive Bayes classifier	96.24%± 1.77	95.17%± 1.11	87.70%± 2.14	82.99%± 1.92
Normal densities based	96.86%± 1.67	96.35%± 1.56	89.88%± 1.37	86.04%± 2.39
Subspace classifier	93.85%± 2.96	93.93%± 1.56	85.52%± 2.82	82.18%± 1.24
Scaled nearest mean	96.26%± 1.22	96.41%± 1.36	89.20%± 1.23	86.35%± 1.37
Nearest mean	83.84%± 2.35	84.23%± 3.02	74.78%±10.72	69.39%±17.02

Different combinations of features are used in normal density based quadratic classifier to discover the effect of these features on classification accuracy and some of the results are summarized in Table 2. When we use degree, clustering coefficient, second connectivity, and contact potential score together, classification accuracy is close to 99%. Even without contact potential score, the method had 98.13% (kCS) prediction accuracy using only the graph properties after outlier analysis. Use of Jernigan contact potentials only decreased the classification accuracy drastically to 51.77%.

Table 2. Classification accuracy rates for different combination of properties with moments. (k: Degree. C: Clustering coefficient. S: Second Connectivity. . J: Profile Score from Jernigan *et. al.*. OA: Outlier Analysis)

	Contact Maps		Delaunay Tes.	
	After OA	Before OA	After OA	Before OA
kCSJ	98.87%± 0.25	98.08%± 0.66	94.81%± 0.60	92.91%± 0.26
CSJ	98.95%± 0.28	97.82%± 0.41	94.60%± 1.18	91.13%± 1.06
SJ	98.15%± 0.25	98.22%± 0.16	89.53%± 0.93	88.36%± 0.48
kC	98.72%± 0.17	97.26%± 0.34	94.72%± 0.32	92.01%± 0.86
k	96.74%± 0.41	96.27%± 0.74	88.68%± 1.21	87.23%± 0.90
kCS	98.13%± 0.60	97.60%± 0.10	94.19%± 1.26	92.12%± 1.17
kS	96.93%± 0.81	95.73%± 0.86	90.43%± 0.74	87.80%± 1.08
J	51.77%± 0.23	48.53%± 0.62	47.71%± 0.84	44.45%± 1.12

Structural properties have more discriminating power, using the degree (k) distribution only we could accurately classify the native and non native structures with 96.74% accuracy. Addition of second connectivity information did not improve the accuracy much. Cliquishness (C) along with degree (k) distribution improved the classification accuracy to 98.72%. Using only the degree and the second connectivity resulted in 96.93% classification accuracy.

6 Conclusion and Discussion

The difference of this study from previous studies can be summarized in four points:

- Using contact maps to derive the structural properties of the proteins yielded much better results than tessellated graphs.
- Combining structural and physicochemical features distinguished the native folds.
- Graph properties have much more discriminative power than the contact potentials.
- Representing the problem as a classification problem, testing the success rate of several classification methods, and building an optimized predictor that can predict native folds about 99 % accuracy.

Classification using the contact potentials only resulted in 51% five fold cross validation accuracy using the quadratic classifier. Thus it is apparent that the structural features are necessary for accurate prediction. As can be seen from the results additional contribution to the prediction accuracy from contact potentials was assumed at less than 1%. Even the non native structures can create favorable interactions between contacting residues so the contact potentials alone are not sufficient to distinguish native structures.

Important structural features were the degree and the clustering coefficient. The second connectivity did not contribute much to the classification accuracy since it is highly correlated to the degree. Previous works focused on the eligibility of different kinds of potentials in discrimination of native folds; this work indicates that structural properties are more important features and, furthermore, these properties can be employed for other problems related to protein structure. This work also shows that contact map provides a better representation of protein structure.

One drawback of our method is all the features that are used in a way capture different aspects of compactness of the protein structure. Our function might fail when trying to identify natively unfolded proteins from random generated counterparts. Since an important feature in the discrimination process is compactness of structure, the method would rule out disordered regions as decoy sets, even though this disorder is a characteristic feature of native states and is functional as well (eg: calcineurin) Such proteins constitute a small subset of all the known protein structures and out of the scope of the proposed work. In addition to this, if decoy sets are generated from naturally unfolded proteins, the native proteins would have more contacts than the artificially generated structures of these native proteins and therefore these naturally unfolded proteins could be captured by our function [23]. This needs to be explored further in a future study.

Another application of our function is to distinguish bad models from good ones (computer generated structures) for protein structure prediction competitions (CASP) [24]. As a preliminary study, we tested the method on CASP VI data set of 59 proteins and 28956 model predictions. Our method correctly assigned 58 proteins as native and 6118 model structures as non native. The predicted non native structures had more than 12 Aº root mean square deviation (rmsd) from the crystal structure. The non native structures assigned as native had much smaller rmsd to the corresponding crystal structures. This shows that the graph properties can easily filter out the bad models. We

are currently working on finding a function using graph properties that can measure closeness of the prediction to the crystal structure on CASP VII data sets and compare it with other ranking methods.

References

1. Baker. D.: Prediction and design of macromolecular structures and interactions. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **361** (2006) 459-463
2. Strogatz S.H.: Exploring Complex Networks. Nature **410** (2001) 268-276
3. Albert. R. and Barabasi. A.-L.: Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74** (2002) 47–97
4. Taylor T. Vaisman I.I.: Graph theoretic properties of networks formed by the Delaunay tessellation of protein structures. *Phys. Rev. E. Stat. Nonlin. Soft. Matter Phys.* **73** (2006) 041925
5. A. R. Atilgan. P. Akan. C. Baysal: Small-World Communication of Residues and Significance for Protein Dynamics. *Biophys. J.* **86** (2004) 85-91
6. Vendruscolo. M.. E. Kussel. and E. Domany: Recovery of Protein Structure from Contact Maps. *Structure Fold. Des.* **2** (1997) 295-306.
7. Fariselli. P. and R. Casadio: A Neural Network Based predictor of Residue Contacts in Proteins. *Protein Eng.* **9** (1996) 941-948.
8. Soyer. A.. J. Chomiller. J.-P. Mornon. R. Jullien. and J.-F. Sadoc: Voronoi Tesselation Reveals the Condensed Matter Character of Folded Proteins. *Phys. Rev. Lett.* **85** (2000) 3532-3535.
9. Vendruscolo. M.. N. V. Dokholyan. E. Paci. and M. Karplus: Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev.* **65** (2002) 061910
10. Miyazawa. S.. and R. L. Jernigan: Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term. for simulation and threading. *J. Mol. Biol.* **256** (1996) 623–644
11. Liang. J. and K.A. Dill: Are proteins Well-Packed? *Biophys. J.* **81** (2001) 751-766
12. Lazaridis. T. and Karplus. M.: Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **10** (2000) 139–145
13. Bonneau. R. and Baker. D.: Ab Initio Protein Structure Prediction: Progress and Prospects. *Annu. Rev. Biophys. Biomol. Struct.* **30** (2001) 173–189
14. McConkey. B.J.. Sobolev. V.. and Eldman. M.: Discrimination of native protein structures using atom–atom contact scoring. *Proc. Natl. Acad. Sci.* **100** (2003) 3215–3220
15. Wang K.. Fain B.. Levitt M.. Samudrala R.: Improved protein structure selection using decoy-dependent discriminatory functions. *BMC Struct. Biol.* **4** (2004) 296
16. Gatchell D. Dennis S. and Vajda S.: Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins* **41** (2000) 518-534
17. Fain B.. Xia H. And Levitt M.: Design of an optimal Chebyshev-expanded discrimination function for globular proteins. *Protein Sci.* **11** (2002) 2010–2021

18. Zhao. Y.. Karypis. G.: Prediction of contact maps using support vector machines. Proceedings of the IEEE Symposium on BioInformatics and BioEngineering. *IEEE Computer Society* (2003) 26- 33
19. Ferdi van der Heijden. Robert P.W. Duin. Dick de Ridder and David M.J. Tax. John Wiley & Sons: Classification. parameter estimation and state estimation - an engineering approach using Matlab. ISBN 0470090138 (2004)
20. G. Wang and R. L. Dunbrack. Jr.: PISCES: a protein sequence culling server. *Bioinformatics* **19** (2003) 1589-1591
21. C. B. Barber. D. P. Dobkin. and H. Huhdanpaa: The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.* **22** (1996) 469-483
22. Vajda S, Jafri MS, Sezerman OU, DeLisi C.: Necessary conditions for avoiding incorrect polypeptide folds in conformational search by energy minimization. *Biopolymers* **33** (1993)173-192
23. Uversky VN, Gillespie JR, and Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **2000**; 41:415-427
24. P.E. Bourne, CASP and CAFASP experiments and their findings, *Methods Biochem Anal* **44** (2003), pp. 501–507.
25. Novotny J, Rashin AA, Bruccoleri RE. 1988. Criteria that discriminate between native proteins and incorrectly folded models. *Proteins Struct. Fund. Genet.* **4**:19-30.
26. B. Park and M. Levitt, Energy Functions that Discriminate X-ray and Near-native Folds from Well-constructed Decoys, *J. Mol. Biol.* **258**, 367, 1996.