JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 23, Number 5, 2016

Pp. 400–411
DOI: 10.1089/cmb.2015.0209

Predicting Designability of Small Proteins from Graph Features of Contact Maps

SUMUDU P. LEELANANDA,[1] ROBERT L. JERNIGAN,[2,3] and ANDRZEJ KLOCZKOWSKI[1,4]

ABSTRACT

Highly designable structures can be distinguished based on certain geometric graphical features of the interactions, confirming the fact that the topology of a protein structure and its residue–residue interaction network are important determinants of its designability. The most designable structures and least designable structures obtained for sets of proteins having the same number of residues are compared. It is shown that the most designable structures predicted by the graph features of the contact diagrams are more densely packed, whereas the poorly designable structures are more open structures or structures that are loosely packed. Interestingly enough, it can also be seen that the highly designable identified are also common structural motifs found in nature.

Key words: contact maps, designability, graph features, interaction network, lattice models, machine learning, network, prediction, structure.

1. INTRODUCTION

Natural proteins are known to fold only to a limited number of folds. Some of these folds are frequently occurring and often referred to as highly designable, whereas some others are rarely observed and are referred to as less designable. Studies have been carried out in the past to understand what makes some protein folds more designable than others and what gives rise to the distribution of designabilities. This concept of protein designability was first introduced by Li et al. (1996). In that lattice model study they defined the designability of a structure as the number of sequences folding to the structure. They found that highly designable protein structures show ''protein-like'' properties. Another interesting aspect of their study is that the structures in the pool differed drastically in their designabilities and the highly designable structures were only a small fraction of all structures. Protein structures are complex systems, and so usually complete enumerations of sequence and structure are not possible. However, with lattice models all conformations can be exactly enumerated. Designability studies do not necessarily require going into atomistic details of structures. Ken Dill demonstrated the utility of lattice models for the study of protein designability (Dill, 1999). That study also showed that despite the simplicity of lattice models, they nonetheless resemble real proteins in many ways.

[1]Nationwide Children's Hospital, Columbus, Ohio.
[2]Iowa State University, Ames, Iowa.
[3]Baker Center for Bioinformatics and Biological Statistics, Ames, Iowa.
[4]The Ohio State University, Columbus, Ohio.

Many studies on designability using lattice models have been reported in the past (Melin et al., 1999; Tang, 2000; Helling et al., 2001; Cejtin et al., 2002; Yang et al., 2007). To simplify the models, two types of residues—hydrophobic and polar (H/P)—are often used with lattice models. Here, the residues in lattice space are characterized as only hydrophobic or polar. All other atomic details of proteins are neglected, and the most important driving force of protein folding is taken to be hydrophobic interactions. The polar side chains are usually directed toward and interact with water, whereas the hydrophobic core of the folded protein consists of nonpolar side chains.

A number of studies have been performed using off-lattice models of proteins as well (Emberly et al., 2002; Miller et al., 2002; Hao-Jun and Yuan-Yuan, 2002). The designability principle not only applies for lattice models of protein folds but also holds for real proteins as well. Wong and Frishman (2006) defined fold designability as the number of families belonging to a particular fold. Interestingly, they also found that many genetic-disease-related proteins have folds that are poorly designable, presumably meaning that these proteins are more susceptible to deleterious conformational changes arising from mutations. In our study we used small real protein structures from the PDB and employed interaction network representations of these structures to predict their designabilities.

We used the interaction networks of proteins and extracted graph theory features from these networks. Network representation of protein structures has been employed in the past in many studies (Kloczkowski and Jernigan, 1997; Dokholyan et al., 2002; Greene and Higman, 2003; Atilgan et al., 2004; Bagler and Sinha, 2005; Brinda and Vishveshwara, 2005; Meyerguz et al., 2007; Milenkovic et al., 2009; Soundararajan et al., 2010; Doncheva et al., 2012; Yan et al., 2014). Krishan et al. (2008) showed the importance, feasibility, and the utility of looking at proteins as networks. Protein systems can be represented as a set of nodes linked by edges (Krishnan et al., 2008). In a study by Doncheva et al. (2012), they used topological network parameters such as connected components, degree of distributions, neighborhood-related parameters, shortest paths, clustering coefficients, and topological coefficients. Brinda and Vishveshwara (2005) represented each amino acid in a protein structure by a node, and the noncovalent interaction strength between two amino acids was considered in the determination of edges. The constructed representations were called protein structure graphs. Sistla et al. (2005) converted the three-dimensional structure defined by the atomic coordinates of proteins into a graph and presented a method for the identification of structural domains of proteins. Jha et al. (2009) showed how topological parameters derived from protein structures can be used for the sequence design for a given set of structures. They used edge-weighted connectivity graphs for ranking residue sites and used optimization techniques to find energy-minimized sequences. They were able to minimize the sequence space for a given target conformation. Lai et al. (2009) used an energy-weighted network of structures in conformation space to study a hydrophobic/hydrophilic model. The energy parameters to weight the vertices were obtained from the Boltzmann factor of each conformation. These parameters represented the importance of each conformation in the conformation space.

It is important to identify structurally and functionally important residues, and binding pockets for drug discovery. However, it is not always possible to find homologs to protein structures in order to make such predictions. Even with a homolog it is still not easy to do this prediction. In work done by Amitai et al. (2004), they were able to identify functional residues of proteins using network analysis. They traced the protein structure in a residue–residue interaction network and used a residue closeness measure in order to predict functionally important residues. The use of graph theory in protein structure studies is discussed in detail in a review by Vishveshwara and coworkers (Patra and Vishveshwara, 2000; Kannan et al., 2001; Vishveshwara et al., 2002).

In general studies of networks, Albert et al. (2000) found that there are highly connected nodes in networks that are crucial for the stability of the network, and these nodes are termed hub-nodes. It is known that real proteins have such crucial residues for stability. Pabuwal and Li (2009) studied these hub-residues specifically for helical membrane and soluble proteins. They concluded that the highly connected amino acid residues in membrane proteins differ from soluble proteins as residues in membrane proteins are exposed to the membrane. They further concluded that the structure–function model of membrane proteins must differ from that of soluble proteins. In a study by Dokholyan et al. (2002) it was shown that topological properties of protein conformations determine their kinetic folding ability. Shakhnovich in his study of designability of conformations found that proteins with larger numbers of residue–residue contacts were more designable (Shakhnovich, 1998).

In our earlier study performed on lattice proteins, all possible compact conformations within a set of 2D and 3D lattice spaces were explored, and we found that complementary interaction graph features can be

used to predict protein designabilities (Leelananda et al., 2011). It was suggested that the topologies of lattice conformations are important determinants of the extent of their designability. Because those findings were encouraging, the same approach was used to address similar questions for real proteins: What makes some protein structures more designable than others? Could interaction graph features be used to answer this question? This study is an extension of our work on lattice models.

2. METHODS

2.1. Selection of datasets

Designability is defined for fixed lengths or a set of structures having the same ''molecular weight.'' It is still an open question of how this can be extended to proteins having different sizes, but this remains a future investigation. Here we utilized a set of conformations having a fixed length. Two sets of data were obtained from the PDB and analyzed. One set consisted of proteins that are all exactly 40 amino acids in length (40-mer set), and the other set consisted of proteins that are all exactly 50 amino acids in length (50-mer set). Because of the high computational cost of calculating designabilities, larger protein sizes were not considered. These sets were further examined manually to carefully remove proteins with missing residues and proteins that have multiple reported occupancies. For NMR structures, only the first of the reported models was considered. The list of proteins used is given in Tables 1a and 1b.

It is important to note that these structures were selected in such a way that they have diversity in the way structural elements are arranged because designability of a structure is measured in relation to all other competing structures. The secondary structure content of these protein chains is shown in Figure 1a and b. The DSSP program was used to identify self-consistently defined secondary structural elements in the datasets (Kabsch and Sander, 1983). There are 8 classes of secondary structure assignments. These 8 classes were contracted into only 3 groups, helix, beta sheet, and coil, for this study as follows: helix (H): H, G, I; sheet (B): E, B; coil (C): S, C, T. The chains were diverse in terms of their secondary structural arrangements.

The pairwise RMSD values were calculated for each set using the CE alignment method. The average RMSD for the two sets of chains were 5.04 and 5.34, respectively, and indicates that the two sets have significant structural diversity. The pairwise variations of the RMSD values for the 40-mer and 50-mer sets of protein chains are shown in Figure 2a and b, respectively. Both average RMSD values and secondary structural content of chains show significant structural diversity in the sets. We used these sets of structures to do the designability calculations.

2.2. Calculating designabilities of structures using binary energy functions

After obtaining the structure sets the sequences were generated. One million random H/P sequences of 40 amino acids and 50 amino acids in length were generated for the 40-mer and 50-mer sets, respectively. Each sequence was threaded on all the C-alpha coarse-grained structures in the structure set (Fig. 3). The contact energies were then calculated using a binary energy function. A contact cutoff distance of 6 Å was used.

Table 1A. PDB IDs of the 45 Proteins Used to Extract the 40-Amino-Acid-Long Chains (40-mer set)

1ADX	1C56	2E3G	1FSB	2NZ3	1ADX	1C56	2E3G	1FSB
1AFO	1D2J	2E5U	1GP8	2RMF	1AFO	1D2J	2E5U	1GP8
1AML	1EDX	2ERL	1HN3	2YSF	1AML	1EDX	2ERL	1HN3
1AOO	1LMM	2GP8	1ICA	2YSG	1AOO	1LMM	2GP8	1ICA
1AQQ	1M7L	2KOE	1JJO	2YSH	1AQQ	1M7L	2KOE	1JJO

Table 1B. PDB IDs of the 36 Proteins Used to Extract the 50-Amino-Acid-Long Chains (50-mer set)

1BK8	1SJU	2CPS	1BK8	1SJU	2CPS	1BK8	1SJU	2CPS
1E8R	1SS3	2DK1	1E8R	1SS3	2DK1	1E8R	1SS3	2DK1
1FDM	1TFI	2EQP	1FDM	1TFI	2EQP	1FDM	1TFI	2EQP
1IFD	1TPM	2FC6	1IFD	1TPM	2FC6	1IFD	1TPM	2FC6

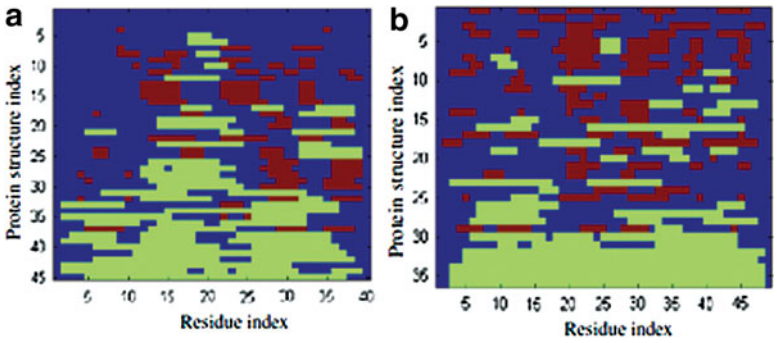

FIG. 1. Secondary structure content: beta sheets (red), alpha helices (green), and coil (blue) for the two protein sets—**(a)** 40-mer set; **(b)** 50-mer set. The two sets are fairly diverse with respect to secondary structural elements.

There are different energy parameters that could be used for the binary alphabet in order to calculate contact energy. We have used the simplest energy parameter set EP1, in which each H-H nonbonded contact interaction is given an energy of −1.0 and all other nonbonded interactions (H-P and P-P) an energy of 0 (arbitrary energy units). This binary energy function has also been used by others (Lau and Dill, 1989; Lipman and Wilbur, 1991). We have also seen that designabilities obtained with EP2 (H-H = −2.3, H-P = −1, and P-P = 0) energy parameter set was comparable with EP1 and designabilities converge even when different energy parameters were used (data not shown). The highly and poorly designable structures obtained in both cases were the same for the two sets of energies as well. The basis for choosing these energies follows from the observation that the most important driving forces for protein folding originate from hydrophobic interactions (Dill, 1999). Hydrophobic residues prefer to be shielded from water, and so they tend to be located inside the core of the protein. Additionally, residues that interact favorably with water (hydrophilic) tend to reside on the surface of the protein in contact with water.

After the threading was carried out for a sequence on all structures in the structure set, the lowest energy structure was identified for this sequence. If a sequence gave the lowest contact energy for two or more structures (degenerate), then that sequence was disregarded (Li et al., 1996). Following this procedure the total number of sequences folding to each structure was obtained, which gives the relative designability of the structure.

2.3. *Generation of contact graphs and the use of graph features*

After obtaining designabilities, residue–residue interaction networks (contact diagrams) of the structures were generated. Residues are different in sizes but a cutoff distance of 6–7 Å (for distances between C^{α} atoms) usually includes most of the closest neighbors. Different cutoff distances have been used in the past. For example, Vendruscolo et al. (2002) used 8.5 Å as their interaction cutoff distance, whereas Atilgan et al. (2004) used 7 Å as theirs. In our study, the contact graphs were generated using a cutoff distance of 6 Å, focusing on the closest set of interactions and the most densely packed part of the structure. First, the coarse-grained alpha carbon representations were obtained for each chain (Fig. 4). The contact diagram was obtained by marking contacts between each C^{α} within the cutoff distance and removing all the bonded interactions.

In these contact diagrams, each graph node represents an amino acid residue and the edges connecting the nodes represent the close contacts between amino acids. Each of these interaction graphs was described using a set of graph features. In other words, the topology of each structure and its interaction network were described using graph features. The graph features used in this analysis were the same features used in our earlier study with the lattice models (Leelananda et al., 2011).

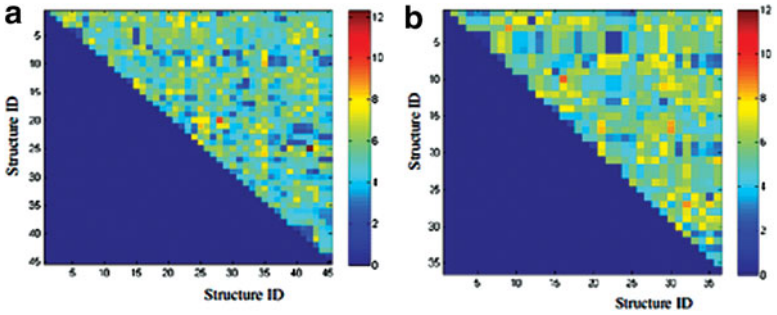

FIG. 2. Pairwise RMSD variations from the CE alignments for the **(a)** 45 structures of the 40-mer set (average RMSD is 5.04) and **(b)** 36 structures of the 50-mer set (average RMSD value is 5.34).

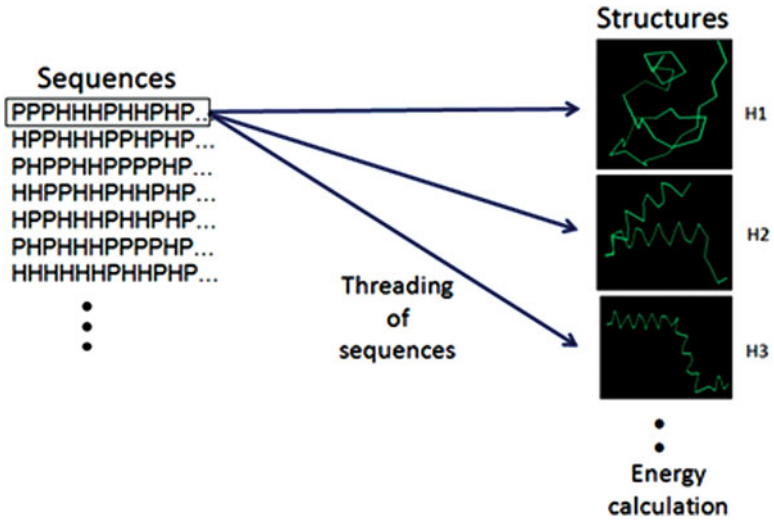

FIG. 3. Schematic diagram showing how the sequences are threaded to calculate contact energies of structures.

Fifteen graph features (or protein conformation interaction graph invariants) were used in this analysis (Fig. 5). These graph features were maximum degree (max_d), average degree (avg_d), maximum shortest path (max_sp), minimum shortest path (min_sp), average shortest path (avg_sp), number of components (compt), number of nodes with minimum degree (n_min_d), number of nodes with maximum degree (n_max_d), number of nodes with average degree (n_avg_d), number of nodes with minimum shortest path (n_min_sp), number of nodes with maximum shortest path (n_max_sp), number of nodes with average shortest path (n_avg_sp), number of nodes with zero degree (zeros), number of nodes with degree one (ones), and number of nodes with degree two (twos). Here, the degree of a node is the number of edges (connections) it has, and the shortest path distance between any two nodes (vertices) is the minimum number of visited edges connecting the two vertices in the interaction graph. The number of components of a graph is the number of maximal connected subgraphs.

2.4. Regression analysis

A numerical value for each of the above features can be found directly from each conformation's interaction graph. Subsequently, a regression curve was obtained for each conformation's designability using the above features. A linear regression curve provides a linear combination of the weighted features that describes the designability of a conformation in terms of the weighted combination of the numerical representation of the graph features. If a nonlinear regression function was used, a slightly better fitting regression function could be obtained. The fit of the regression function was calculated based on the correlation of its output with the actual number of sequences that fold onto the conformation being examined. Regression analysis was carried out using the Weka software (Hall et al., 2009). Regression functions were constructed using all of the features and taking each feature individually. Going further, a designability range for which each structure folded to was predicted instead of predicting the exact designability using linear regression. Better correlations were

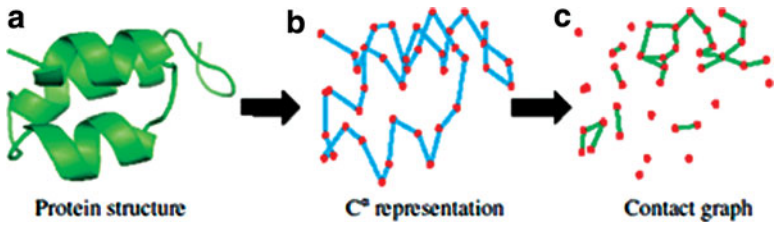

FIG. 4. Schematic description of the method employed to generate the contact graph: **(a)** the protein structure used, **(b)** C^{α} representation of the protein structure derived by connecting all consecutive C^{α} atoms, and **(c)** the contact graph obtained from **(b)** by marking contacts between each C^{α} within the cutoff distance (6 Å) and removing all the bonded interactions. The contact graphs or networks are described using numerical values for graph features.

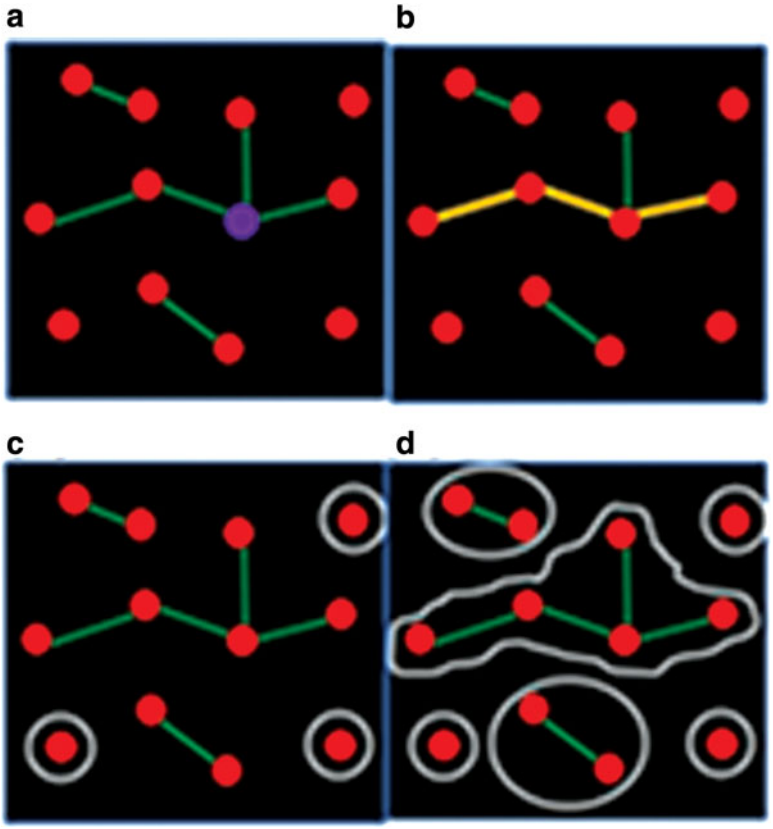

FIG. 5. Examples of a few graph features using a simple contact map: **(a)** maximum degree of connectivity = 3; **(b)** the maximum shortest path distance = 3; **(c)** the number of nodes with degree zero = 3; **(d)** the number of separate compartments = 6.

obtained when ranges were predicted instead of the exact numerical designabilities as shown in our previous study using lattice models (Leelananda et al., 2011).

The number of sequences folding to a particular conformation is given by N_s, and this is also designated as the designability of that structure. First, the distribution of designabilities for all the possible conformations for a particular model was obtained. A naïve Bayes (NB) classifier was then used to see if the features describing each fold could be used to predict its designability range. In order to do this, the designability distribution was first discretized into three bins using the Weka software (Figs. 6 and 7) such that the overall distribution of designability was preserved. This process of binning simplifies the calculations. We have also obtained results for larger numbers of discretized bins and saw that results were comparable with those obtained by using just three bins. In the case of the 50-mer set, there were two extremely highly designable structures that stood out from the rest of the structures. Therefore, logarithms of the values of the designability, instead of the designability values themselves, were used in order to obtain better binning. Machine learning algorithms (NB) were then used to find the range of designability of a structure by using graph features describing it, and if the actual range fell within the range predicted, then the prediction was considered correct. Ten-fold cross-validation of data was used for predictions.

2.5. Naïve Bayes prediction

Bayes' theorem states that given a hypothesis h and data D that bears on the hypothesis,

$$P(h|D) = [P(D|h) \times P(h)]/P(D) \quad (1)$$

where P(h) is independent probability of h; P(D), independent probability of D; P(D|h), conditional probability of D for given h; and P(h|D), conditional probability of h for given D.

An NB classifier is a simple probabilistic classifier based on Bayes's theorem with the independence assumption. In other words, such a classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. In the training step, for each conformation described by 15 vectors or features, $P(\text{feature}_i|\text{range}_j)$, where $1 \le i \le 15$ and $1 \le j \le 3$ for the

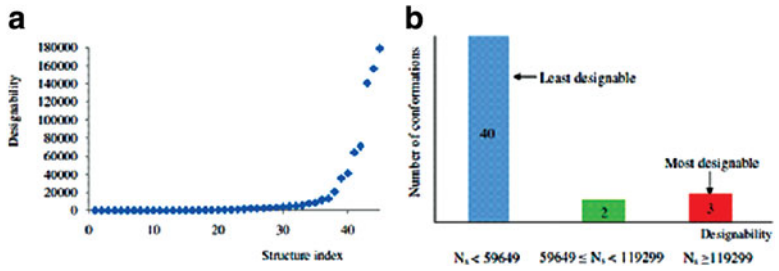

FIG. 6. Discretizing the designability (Ns) distribution for the 40-mer set into three bins: **(a)** the designability distribution; **(b)** discretized distribution using the Weka software (red, most designable; green, intermediate; blue, least designable, with designability ranges shown along the bottom). The number of structures in each bin is shown inside the colored bars.

three selected bins, was calculated along with $P(range_j)$ and $P(feature_i)$. In the testing step, the $P(range_j|feature_i)$ was calculated using Bayes's theorem. This way all of the features that define a conformation can be used together to predict the most probable range for its designability. A range for the designability value or a confidence interval was predicted for each structure, and the prediction was considered ''correct'' if the actual designability value lies in that range of maximum probability. The total energy of all sequences folding to each structure was also calculated and averaged over all the sequences folding to the structure to get the average energy of sequences folding to the structure (note that the energies are negative values). In order to calculate the contact density of each structure, the number of nonbonded contacts of each structure is found and averaged over the chain length.

3. RESULTS

Figures 6 and 7 show the discretization of the designability distribution into bins. For the 40-mer set, more structures have designabilities of less than 20,000 in contrast to only a few structures having designabilities of more than 10,000. These structures are identified as the most designable and have total designabilities greater than 140,000. For the 50-mer set, clearly two structures stand out for their designabilities. Highly designable structures show higher average energies (Fig. 8). These structures also have higher contact densities (Fig. 9). There are some poorly designable structures that also show high contact densities and high energies. However, highly designable structures always have higher average energies and contact densities.

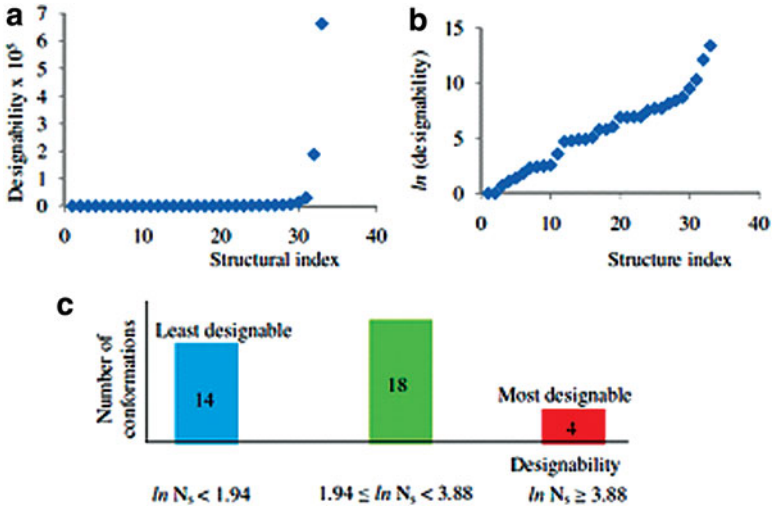

FIG. 7. Discretizing the designability (N_s) distribution of the 50-mer set into three bins: **(a)** the designability distribution; **(b)** the distribution of the logarithm of designability; **(c)** the discretized designability distribution of **(b)** using the Weka software (red, most designable; green, intermediate; blue, least designable, where the designability ranges are shown at the bottom). The number of structures in each bin is shown inside the bin.

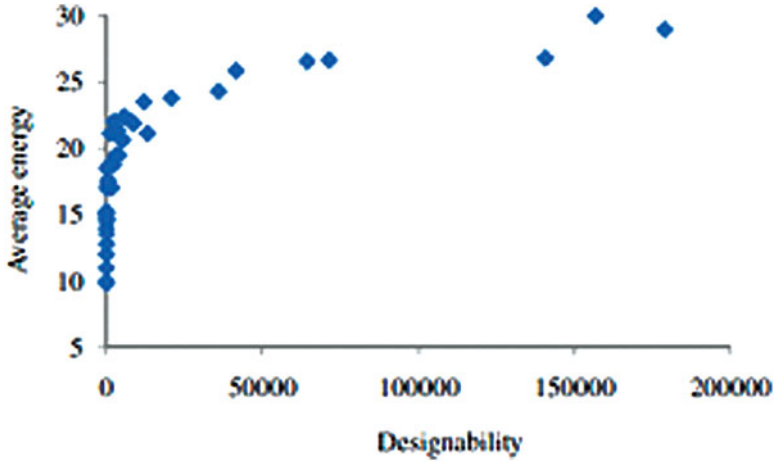

FIG. 8. The relationship between designability and average energy for the 40-mer set. Here, it is seen that the highly designable structures appear to be energetically more favorable. (Note: Negative energies plotted, so high values are favored.)

A positive correlation between the topological arrangement of conformation and its designability is observed from linear regression analysis. Linear regression analysis for the 40-mer set gives a correlation coefficient of 0.70. The relationship that gives the best fit to designability is as follows:

Designability = $[2\times10^3 \times$ (maximum degree)$]$ + $[4\times10^3 \times$ (average degree)$]$ − $[3\times10^5]$

The correlation coefficient for linear regression for the 50-mer set is 0.85 but the best-fit equation for this case is more complex and more features are involved. When NB 10-fold cross-validation is used for predictions, a prediction accuracy of 93% (AUC = 0.86) is obtained for the 40-mer set. The prediction accuracy for the 50-mer set is 59.3% (AUC = 0.62). The corresponding AUC values for the 3 ranges—the lowest designable, intermediately designable, and the most designable—are shown in Table 2. The highly designable range prediction AUC values are 0.92 for both the 40-mer and the 50-mer set (Table 2). The least designable range is predicted with AUCs of 0.89 for the 40-mer set and 0.62 for the 50-mer set. As expected, the intermediate range is not well distinguishable from highly and poorly designable structures. However, the two extreme ranges are sharply distinguishable.

According to the regression analysis, the most important feature in predicting the designability of the 40-mer set is the maximum degree and the average degree of connectivity of structure nodes. For the 50-mer set, in addition to the maximum degree of connectivity and the average degree of connectivity, the number with the average shortest path and the number with the maximum shortest path are also found to be important for predicting designability.

The most designable structures obtained are found to be popular structural motifs found in nature (Figs. 10a and 11a). One of the motifs for the 40-mer set is a helix-loop-helix motif and the other is a beta-hairpin-loop helix–like motif. The least designable structures are more extended-type structures (Figs. 10b and 11b). For the 50-mer set the most designable structures are an up-down helix bundle structure and a ribbon-like structure. Highly designable structures always show higher maximum degrees and higher average degrees of connectivity (Fig. 12a and b). More designable structures always have higher values for these two degree measures, but not exclusively so. There are some poorly designable structures with high values.

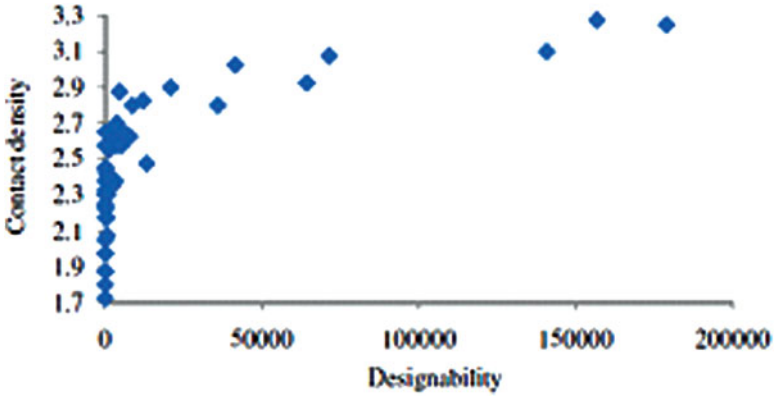

FIG. 9. Relationship between the designability and contact density for the 40-mer set. Highly designable structures have higher contact densities.

Table 2. AUC Values of Prediction of Designability Ranges for the 40-mer and 50-mer Sets

Designability range	*40-mer set*	*50-mer set*
Highly designable	0.92	0.92
Intermediate	0.21	0.55
Poorly designable	0.89	0.62

The highly designable structures in both sets are predicted with high AUC.

4. DISCUSSION

The graphical features that describe the topology of a protein structure are important determinants of its designability. Most designable and least designable structures can clearly be distinguished based on certain interaction network features. The most important features in predicting the designability for the 40-mer set were found to be the maximum degree and the average degree of connectivity of structure nodes, as could be seen in the regression equation. However, for the 50-mer set the regression analysis was more complex and more features were need for the prediction of the designabilities. As the size of the protein sets increases, the predictions require more structural features. This is true for both regression analysis and NB 10-fold cross validation. Further analysis of the predicted highly and poorly designable structures showed interesting properties of these structures.

Highly designable structures were found to be energetically more favorable and more stable than poorly designable structures. It is expected that frequently occurring structures should be more stable than other structures. These highly designable structures were also more densely packed structures and have more interactions. Highly designable structures always had higher maximum degrees and higher average degrees of connectivity. More designable structures always have higher values for these two measures, but not exclusively so. There are also some poorly designable structures with high values for connectivity

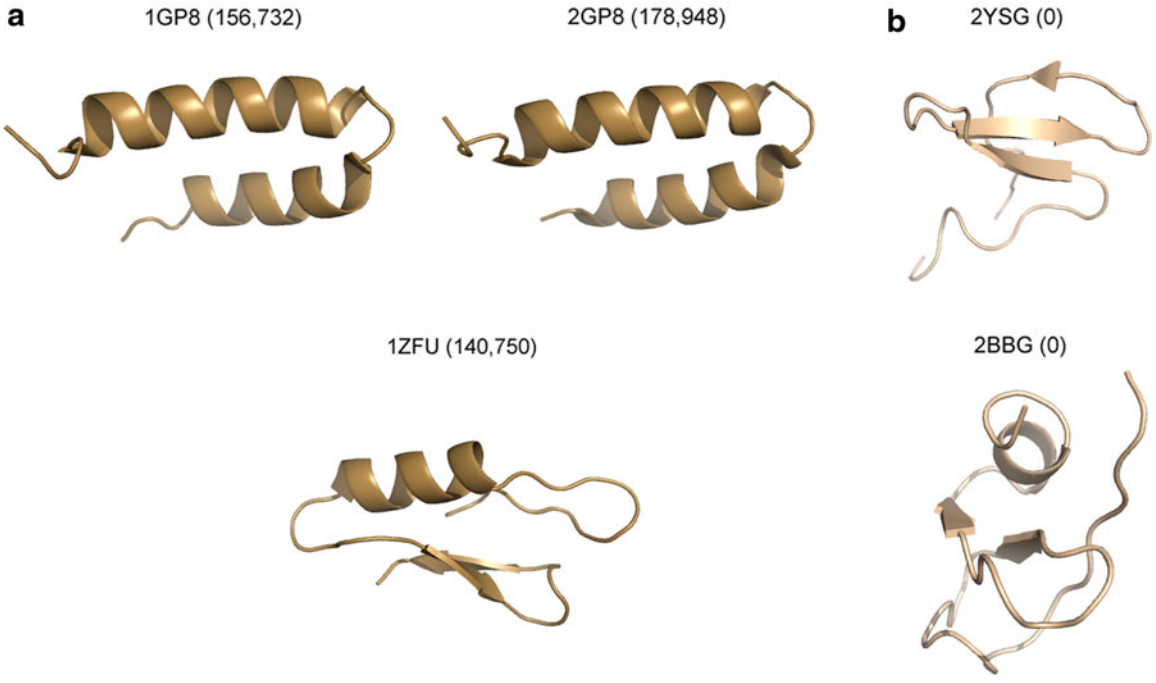

FIG. 10. The most designable and least designable chains for the 40-mer set: **(a)** Most designable: helix-loop-helix motifs 1GP8 and 2GP8, and beta-hairpin-loop-helix motif 1ZFU. These structures are popular structural motifs found in nature. The number of sequences folding to each of these structures is shown within parentheses. **(b)** Least designable: 2BBG and 2YAS. These structures are more extended and have more open types of structures.

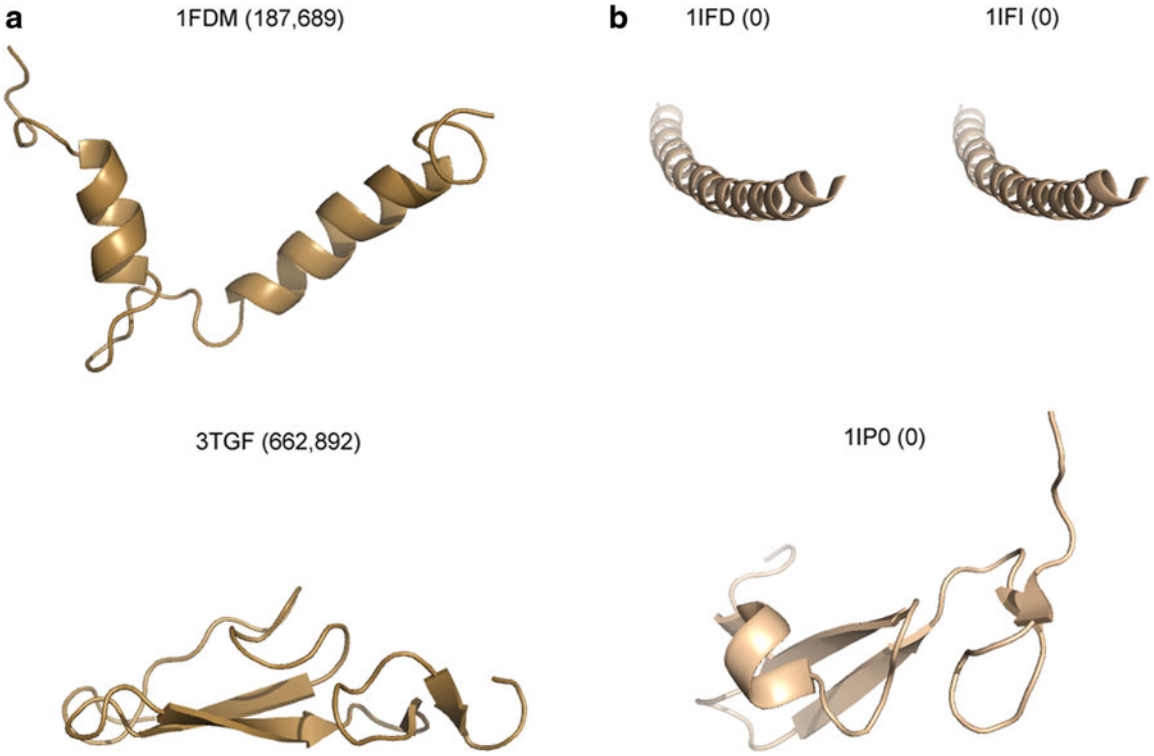

FIG. 11. The most designable and least designable chains for the 50-mer set. **(a)** The most designable motifs 1FDM and 3TGF. These structures take an up-down helix bundle and a ribbon-like structure, which are common structural motifs. **(b)** The least designable motifs 1IFD, 1IFI, and 1IP. Two of these structures are extended single alpha helical structures, whereas the other is a more distorted ribbon structure.

measures. The contact densities of structures showed that there is clearly a distinction between highly and poorly designable structures. These results also agree with England et al. (2003), who compared thermophilic and mesophilic protein analogs and found that, based on the contact densities of these proteins, these functional analogs could be distinguished.

Interestingly, highly designable structures obtained were also popular and more abundant structural motifs found in nature for both the 40-mer and the 50-mer cases (Figs. 10 and 11). Recurring motifs in

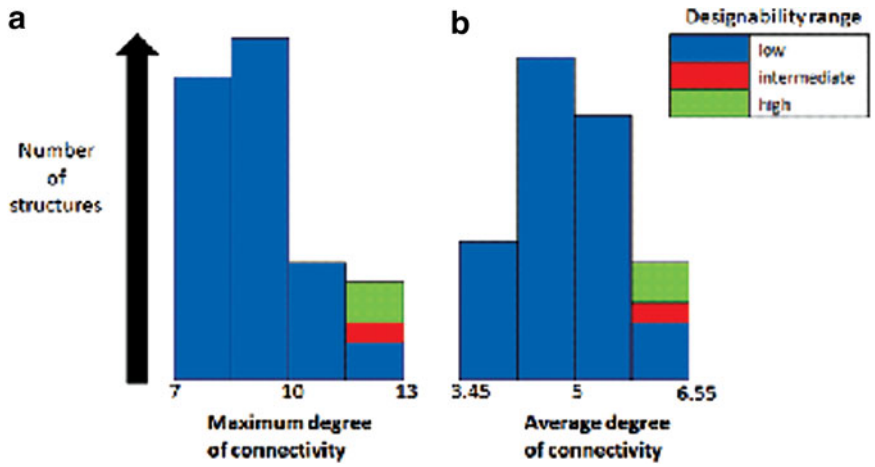

FIG. 12. Number of cases falling into each of the **(a)** maximum degree bin and **(b)** average degree bin (red, intermediate; light blue, most designable; dark blue, poorly designable). Highly designable structures always have higher maximum degrees and higher average degrees of connectivity. More designable structures always have higher values for these two measures, but not exclusively so. There are also some poorly designable structures with high values.

nature must be able to accommodate a wide range of sequences and our results are in good agreement with this theory. On the other hand, least designable structures are more extended and open structures with more loop segments. Although these structures certainly often occur, they are not expected to be highly designable or frequently occurring.

Finding what makes some protein folds more designable than others has many implications ranging from computer-aided drug design to predicting properties of proteins of unknown function. Graph features may be used to pick out these most designable motifs and to sample structure space as well. Graph features may also be used in protein design or for the inverse protein folding problem to identify the compatible sequences that can fold to a particular structure of interest. Algorithms can be developed to satisfy the feature constraints and design particular structures.

There is a high demand for improving protein structure prediction methods because the gap between the number of experimentally solved protein structures and the number of known sequences continues to accelerate. The knowledge of protein structure is also critical for comprehending their function. Also, the design of completely new proteins with desired properties that have not yet been found in nature is becoming increasingly important. In the last decade we have witnessed the rise of synthetic biology, including *de novo* design of proteins that were first theoretically conceived and were then synthesized. Our work has implications that can be used as inputs in computer-aided design of protein drugs and protein structure prediction. The important graph features we obtained could be used to sample structure space and algorithms could be developed to satisfy the feature constraints. These could then be used in designing structures of interest and for identifying the compatible sequences that can fold into a particular structure.

ACKNOWLEDGMENTS

This work was supported by the NSF Grant MCB-1021785 and National Institutes of Health Grants R01GM081680 and R01GM072014.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Albert, R., Jeong, H., and Barabasi, A.L. 2000. Error and attack tolerance of complex networks. *Nature* 406, 378–382.

Amitai, G., Shemesh, A., Sitbon, E., et al. 2004. Network analysis of protein structures identifies functional residues. *J. Mol. Biol.* 344, 1135–1146.

Atilgan, A.R., Akan, P., and Baysal, C. 2004. Small-world communication of residues and significance for protein dynamics. *Biophys. J.* 86, 85–91.

Bagler, G., and Sinha, S. 2005. Network properties of protein structures. Phys. A Stat. Mech. Appl. 346, 1–2.

Brinda, K.V., and Vishveshwara, S. 2005. A network representation of protein structures: Implications for protein stability. *Biophys. J.* 89, 4159–4170.

Cejtin, H., Edler, J., Gottlieb, A., et al. 2002. Fast tree search for enumeration of a lattice model of protein folding. *J. Chem. Phys.* 116, 352–359.

Dill, K.A. 1999. Polymer principles and protein folding. *Protein Sci.* 8, 1166–1180.

Dokholyan, N.V., Li, L., Ding, F., et al. 2002. Topological determinants of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* 99, 8637–8641.

Doncheva, N.T., Assenov, Y., Domingues, F.S., et al. 2012. Topological analysis and interactive visualization of biological networks and protein structures. *Nat. Protocols* 7, 670–685.

Emberly, E.G., Miller, J., Zeng, C., et al. 2002. Identifying proteins of high designability via surface-exposure patterns. *Proteins Struct. Funct. Bioinform.* 47, 295–304.

England, J.L., Shakhnovich, B.E., and Eugene, I.S. 2003. Natural selection of more designable folds: A mechanism for thermophilic adaptation. *Proc. Natl. Acad. Sci. U. S. A.* 100, 28727–28731.

Greene, L.H., and Higman, V.A. 2003. Uncovering network systems within protein structures. *J. Mol. Biol.* 334, 781–791.

Hall, M., Frank, E., Holmes, G., et al. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newslett.* 11, 10–18.

Hao-Jun, L., and Yuan-Yuan, W. 2002. Influence of monomer types on the designability of a protein-model chain. *Chin. Phys. Lett.* 19, 1382.

Helling, R., Li, H., Melin, R., et al. 2001. The designability of protein structures. *J. Mol. Graph. Model.* 19, 157–167.

Jha, A.N., Ananthasuresh, G.K., and Vishveshwara, S. 2009. A search for energy minimized sequences of proteins. *PLoS ONE* 4, e6684.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637.

Kannan, N., Selvaraj, S., Gromiha, M.M., et al. 2001. Clusters in alpha/beta barrel proteins: Implications for protein structure, function, and folding: A graph theoretical approach. *Proteins Struct. Funct. Bioinform.* 43, 103–112.

Kloczkowski, A., and Jernigan, R.L. 1997. Efficient method to count and generate compact protein lattice conformations. *Macromolecules* 30, 6691–6694.

Krishnan, A., Zbilut, J.P., Tomita, M., et al. 2008. Proteins as networks: Usefulness of graph theory in protein science. *Curr. Protein Peptide Sci.* 9, 28–38.

Lai, Z., Su, J., Chen, W., et al. 2009. Uncovering the properties of energy-weighted conformation space networks with a hydrophobic-hydrophilic model. *Int. J. Mol. Sci.* 10, 1808–1823.

Lau, K.F., and Dill, K.A. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* 22, 3986–3997.

Leelananda, S.P., Towfic, F., Jernigan, R.L., et al. 2011. Exploration of the relationship between topology and designability of conformations. *J. Chem. Phy.* 134, 235101.

Li, H., Helling, R., Tang, C., et al. 1996. Emergence of preferred structures in a simple model of protein folding. *Science* 273, 666–669.

Lipman, D.J., and Wilbur, W.J. 1991. Modeling neutral and selective evolution of protein folding. *Proc. Biol. Sci.* 245, 7–11.

Melin, R., Li, H., Wingreen, N.S., et al. 1999. Designability, thermodynamic stability, and dynamics in protein folding: A lattice model study. *J. Chem. Phys.* 110, 1252–1262.

Meyerguz, L., Kleinberg, J., and Elber, R. 2007. The network of sequence flow between protein structures. *Proc. Natl. Acad. Sci. U. S. A.* 104, 11627–11632.

Milenkovic, T., Filippis, I., Lappe, M., et al. 2009. Optimized null model for protein structure networks. *PLoS ONE* 4, e5967.

Miller, J., Zeng, C., Wingreen, N.S., et al. 2002. Emergence of highly designable protein-backbone conformations in an off-lattice model. *Proteins Struct. Funct. Bioinform.* 47, 506–512.

Pabuwal, V., and Li, Z. 2009. Comparative analysis of the packing topology of structurally important residues in helical membrane and soluble proteins. *Protein Eng. Des. Sel.* 22, 67–73.

Patra, S.M., and Vishveshwara, S. 2000. Backbone cluster identification in proteins by a graph theoretical method. *Biophys. Chem.* 84, 13–25.

Shakhnovich, E.I. 1998. Protein design: A perspective from simple tractable models. *Fold. Des.* 3, R45–R58.

Sistla, R.K., Brinda, K.V., and Vishveshwara, S. 2005. Identification of domains and domain interface residues in multidomain proteins from graph spectral method. *Proteins Struct. Funct. Bioinform.* 59, 616–626.

Soundararajan, V., Raman, R., Raguram, S., et al. 2010. Atomic interaction networks in the core of protein domains and their native folds. *PLoS ONE* 5, e9391.

Tang, C. 2000. Simple models of the protein folding problem. *Phys. A Stat. Mech. Appl.* 288, 1–4.

Vendruscolo, M., Dokholyan, N.V., Paci, E., et al. 2002. Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E* 65, 061910.

Vishveshwara, S., Brinda, K.V., and Kannan, N. 2002. Protein structure: Insights from graph theory. *J. Theor. Comput. Chem.* 1, 187–211.

Wong, P., and Frishman, D. 2006. Fold designability, distribution, and disease. *PLoS Comput. Biol.* 2, e40.

Yan, W., Sun, M., Hu, G., et al. 2014. Amino acid contact energy networks impact protein structure and evolution. *J. Theor. Biol.* 355, 95–104.

Yang, J.Y., Yu, Z.G., and Anh, V. 2007. Correlations between designability and various structural characteristics of protein lattice models. *J. Chem. Phys.* 126, 195101.

Address correspondence to:
Prof. Andrzej Kloczkowski
Nationwide Children's Hospital
Battelle Center for Mathematical Medicine
575 Children's Crossroad
Columbus, OH 43215

E-mail: andrzej.kloczkowski@nationwidechildrens.org