2011

Exploration of the relationship between topology and designability of conformations

Sumudu P. Leelananda
Iowa State University

Fadi Towfic
Iowa State University

Robert L. Jernigan
Iowa State University, jernigan@iastate.edu

Andrzej Kloczkowski
Iowa State University

Exploration of the relationship between topology and designability of conformations

Abstract

Protein structures are evolutionarily more conserved than sequences, and sequences with very low sequence identity frequently share the same fold. This leads to the concept of protein *designability*. Some folds are more designable and lots of sequences can assume that fold. Elucidating the relationship between protein sequence and the three-dimensional (3D) structure that the sequence folds into is an important problem in computational structural biology. Lattice models have been utilized in numerous studies to model protein folds and predict the designability of certain folds. In this study, all possible compact conformations within a set of two-dimensional and 3D lattice spaces are explored. Complementary interaction graphs are then generated for each conformation and are described using a set of graph features. The full HP sequence space for each lattice model is generated and contact energies are calculated by threading each sequence onto all the possible conformations. Unique conformation giving minimum energy is identified for each sequence and the number of sequences folding to each conformation (designability) is obtained. Machine learning algorithms are used to predict the designability of each conformation. We find that the highly designable structures can be distinguished from other non-designable conformations based on certain graphical geometric features of the interactions. This finding confirms the fact that the topology of a conformation is an important determinant of the extent of its designability and suggests that the interactions themselves are important for determining the designability.

Disciplines

Biochemistry | Bioinformatics | Biophysics | Molecular Biology | Structural Biology

Comments

This article is published as Leelananda, Sumudu P., Fadi Towfic, Robert L. Jernigan, and Andrzej Kloczkowski. "Exploration of the relationship between topology and designability of conformations." The Journal of chemical physics 134, 235101 (2011). doi: 10.1063/1.3596947. Posted with permission.

Exploration of the relationship between topology and designability of conformations

Sumudu P. Leelananda, Fadi Towfic, Robert L. Jernigan, and Andrzej Kloczkowski

Citation: The Journal of Chemical Physics **134**, 235101 (2011);
View online: https://doi.org/10.1063/1.3596947
View Table of Contents: http://aip.scitation.org/toc/jcp/134/23
Published by the American Institute of Physics

Exploration of the relationship between topology and designability of conformations

Sumudu P. Leelananda,[1,2] Fadi Towfic,[1,3] Robert L. Jernigan,[1,2] and Andrzej Kloczkowski[1,2,4,5]
[1]*L. H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, Iowa 50010, USA*
[2]*Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, Iowa 50010, USA*
[3]*Department of Computer Science, Iowa State University, Ames, Iowa 50010, USA*
[4]*Battelle Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital, Columbus, Ohio 43205, USA*
[5]*Department of Pediatrics, The Ohio State University College of Medicine, Columbus, Ohio 43205, USA*

(Received 1 October 2010; accepted 16 May 2011; published online 15 June 2011)

Protein structures are evolutionarily more conserved than sequences, and sequences with very low sequence identity frequently share the same fold. This leads to the concept of protein *designability*. Some folds are more designable and lots of sequences can assume that fold. Elucidating the relationship between protein sequence and the three-dimensional (3D) structure that the sequence folds into is an important problem in computational structural biology. Lattice models have been utilized in numerous studies to model protein folds and predict the designability of certain folds. In this study, all possible compact conformations within a set of two-dimensional and 3D lattice spaces are explored. Complementary interaction graphs are then generated for each conformation and are described using a set of graph features. The full HP sequence space for each lattice model is generated and contact energies are calculated by threading each sequence onto all the possible conformations. Unique conformation giving minimum energy is identified for each sequence and the number of sequences folding to each conformation (designability) is obtained. Machine learning algorithms are used to predict the designability of each conformation. We find that the highly designable structures can be distinguished from other non-designable conformations based on certain graphical geometric features of the interactions. This finding confirms the fact that the topology of a conformation is an important determinant of the extent of its designability and suggests that the interactions themselves are important for determining the designability. © *2011 American Institute of Physics*. [doi:10.1063/1.3596947]

I. INTRODUCTION

Understanding the physical characteristics responsible for the folding of protein sequences to their native structures is one of the most important problems in computational structural biology because it requires a deeper understanding of the protein *sequence-structure* relationship. This is, however, an extremely challenging problem. Even though there are tens of thousands of protein structures in the Protein Data Bank (PDB),[1] the structures they take are limited to some thousands of folds.[2,3] It is well known that protein structures are evolutionarily more conserved than sequences[4] and often sequences that have low sequence identity can share the same fold.[5] This leads to the concept of protein *designability*.[6] Designability of a particular conformation is the number of different sequences that folds to the conformation giving unique minimum energy. As noted by Li *et al.*, some folds that are more designable will have many more primary amino acid sequences mapping to that same tertiary structural fold.[6]

One of the most straightforward approaches for elucidating the relationship between sequence and fold structure is to utilize lattices so that all protein conformations can be exactly enumerated. It has been well established that despite their simplicity, such models can resemble real proteins in many ways.[6] Numerous studies have been conducted on lattice models to understand protein designability.[6–11] Studies have also been done on off-lattice models[12] and semi-off-lattice models of proteins.[13] Emberly *et al.* used off-lattice models of proteins and found that the surface exposure pattern of folded structures is related to their designability.[14]

Network representations of protein structures have been employed in the past.[15–19] Brinda *et al.* 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.[20] The constructed representations were called protein structure graphs (PSGs). Sistla *et al.* converted the three-dimensional (3D) structure, defined by the atomic coordinates of proteins into a graph and presented a method for the identification of structural domains of proteins.[21] Jha *et al.* showed how topological parameters derived from protein structures can be used for the sequence design for a given set of structures.[22] They used edge weighted connectivity graph for ranking the residue sites and used optimization techniques to find energy minimizing sequences. By

this way, they were able to minimize the sequence space for a given target conformation. The use of graph theory in protein structures is discussed in detail in a review by Vishveshwara *et al*.[23]

The designability principle that holds for simple lattice models of protein folds holds for real proteins as well. Wong *et al*. defined fold designability as the number of families belonging to a particular fold.[5] Interestingly, they also found that many genetic-disease related proteins have folds that are poorly designable, meaning presumably that these proteins are more susceptible to conformational changes arising from mutations. Full enumeration of an entire fold-space (conformation space) for a specific lattice model (e.g., $3 \times 3 \times 3$) allows us to address designability and to directly answer questions regarding the significance of the relationship between protein sequences and the possible folds such sequences may take in three-dimensional (3D) space.

Since the role of the interaction parameters is not important for choosing highly designable structures (see the Methods section), geometry is thought to be an important factor for determining the designability of a conformation.[24] England *et al.* observed that a fold's tertiary topology correlates with the fold's designability.[25] Hoang *et al.* demonstrated that native protein folds can arise from the considerations of symmetry and geometry of their polypeptide chains using a simple physical model.[26] They further showed that the limited number of protein folds can arise from the geometrical constraints that are imposed by the steric interactions and hydrogen bond interactions. Banavar *et al*. suggested that symmetry and geometry constraints lead to a finite number of protein folds similar to the way they impose constraints on the limited number of types of infinite crystal lattice structures.[27] To further investigate the role of a fold's topology and its designability, it is important to understand exactly what features of the topology of a particular conformation affect its designability. To address this issue, we utilize protein structure graphs to represent lattice models and investigate the relationship between various graph features based on the structure graphs and designable conformations. We learn that there are several graph features that aid in the prediction of the extent of designability and that by using these features the most designable conformations can be distinguished from the rest of structures.

II. METHODS

In order to explore the full conformation space for each lattice model, all possible compact conformations within 2D lattices—the 3×4, 4×4, and 5×5 and 3D lattices—$2 \times 2 \times 3$ and $3 \times 3 \times 3$ are enumerated. Hamiltonian walks are utilized, where all sites are visited once and only once (excluded volume condition holds), and empty unvisited sites (vacancies) are not allowed. Enumerations for some of these models have been carried out in the past.[6,9,28–31]

Each of the aforementioned lattice models represents proteins having different numbers of residues (see Table I). The total number of possible walks (conformations) without including rotational and reflection symmetries are shown in Table I. We have also shown the results for the hexagonal and triangular lattice models studied by Peto *et al*.[32,33] When the size (number of nodes) of the models increases, the number of possible conformations increases exponentially.

TABLE I. Lattice models used and corresponding numbers of conformations and H/P sequences.

Lattice model	Number of conformations	Number of H/P sequences
3×4	31	4096
4×4	69	65536
5×5	1081	33,554,432
2×2×3	73	4096
3×3×3	103,346	134,217,728
Hexagonal	22,104	524,288
Triangular	20,843	2,097,152

For example, a 3×4 model represents a protein with 12 residues. The total number of possible walks (conformations) is 31 without including rotational and reflection symmetries and if head-tail symmetrical conformations are excluded. We use a binary hydrophobic-polar (H/P) model to generate all possible amino acids sequences for each lattice model. For the 3×4 case this amounts to a total of 2^{12} (4096) different HP sequences having two distinguishable ends; the C-terminal end and the N-terminal end. For larger models such as the $3 \times 3 \times 3$, 5×5 random sampling of sequences has been employed.

Generated sequences can be threaded onto the enumerated conformations and an energy function may be used to calculate the energy of each threading.

There are many energy functions that could be utilized for the binary alphabet, and in this paper we use a simple function where each H–H non-bonded contact interaction is given an energy of -1.0 and all other non-bonded interactions (H–P and P–P) are given energy 0 in arbitrary energy units.[34–36] The reasoning for choosing this energy function is the belief that the most important driving force of protein folding is *hydrophobic* interactions.[37] Hydrophobic residues prefer to be shielded from water, 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.

Interaction parameters used to calculate energies of conformations need only to have basic physical features: (i) The condition $E_{PP} \geq E_{HP} > E_{HH}$ that reflects the protein feature that hydrophobic residues are hidden inside the core. (ii) The condition $2E_{HP} > E_{HH} + E_{PP}$ corresponds to the tendency of the mixture of the H and P residues to segregate. The detailed numerical values of energy parameters are less important. Additionally it has been shown by Li *et al.* who used the full 20-letter amino acid alphabet and corresponding Miyazawa-Jernigan matrix of contact interactions, and studied protein designability for the $3 \times 3 \times 3$ cube model, that the results are in good qualitative agreement with those obtained earlier for a simple H/P model.[38]

In order to compute the designability of a specific fold, we find the total number of different sequences that folds to each conformation with the lowest non-degenerate energy. All possible HP sequences are generated and each sequence is threaded onto all of the conformations, after which, the

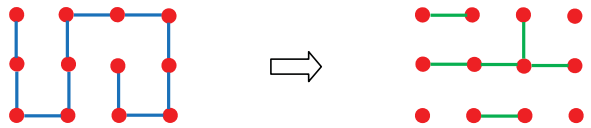

FIG. 1. Interaction graph (right) complementary to one conformation of the 3 × 4 lattice (left).

TABLE II. Correlation coefficients for non-linear regression analysis for the training set and for 10-fold cross-validation.

Lattice model	Correlation coefficients for training set	Correlation coefficient 10-fold cross-validation
3×4	0.65	0.42
4×4	0.60	0.46
5×5	0.66	0.53
2×2×3	0.55	0.44
3×3×3	0.57	0.50

contact energies are calculated. The conformation that has the lowest energy for the particular threaded sequence is identified from the conformation space (lowest energy conformation is assumed to be the conformation that the sequence 'folds' to). When there exist more than one conformation giving the same minimum energy for a particular sequence (degeneracy), then that sequence is disregarded. In other words, we only consider non-degenerate (unique) lowest energy giving sequences (see supplementary Table S2).[39] This energy calculation is repeated for all the sequences in the sequence space. Thus, all the sequences in the sequence space are tested against all the structures in the structure space in order to identify the structures these sequences adopt. As such, we calculate the number of sequences that fold to each conformation while maintaining a unique ground state energy. There are some conformations to which many sequences fold, and such conformations are called highly designable structures. There are other conformations to which none or only a small number of sequences fold, such conformations are deemed to be poorly designable.

Once the folds for a given shape are computed (e.g., for 3 × 4 lattice), we use the lattice conformation to generate a corresponding complementary nearest neighbor non-bonded interaction graph (hereafter referred to as an interaction graph for simplicity). Consider as an example the lattice conformation shown in Fig. 1. The interaction graph is generated by drawing horizontal or vertical non-diagonal edges between nodes that have a Euclidean distance of exactly 1 in the unit lattice that are not already connected by edges and by dropping the edges between nodes that already exist in the lattice.

We consider topological features that can be used to 'define' a conformation based on the interaction graphs of that lattice conformation. The graph features (or graph invariants) that we have used in our analysis are

(i) maximum degree (max_d),
(ii) average degree (avg_d),
(iii) minimum shortest path (min_sp),
(iv) maximum shortest path (max_sp),
(v) average shortest path (avg_sp),
(vi) number of components (compt),
(vii) number of nodes with minimum degree (n_min_d),
(viii) number of nodes with maximum degree (n_max_d),
(ix) number of nodes with average degree (n_avg_d),
(x) number of nodes with minimum shortest path (n_min_sp),
(xi) number of nodes with maximum shortest path (n_max_sp),
(xii) number of nodes with average shortest path (n_avg_sp),
(xiii) number of nodes with zero degree (zeros),
(xiv) number of nodes with degree one (ones), and
(xv) 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. Number of components of a graph is the number of maximal connected subgraphs.

A numerical value for each of the above features can be found directly from each conformation's interaction graph. Subsequently, a regression curve may be 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 relation to the weighted combination of the numerical representation of the graph features. If a non-linear regression function is utilized, a slightly better fitting regression function can be obtained (the fit of the regression function is calculated based on the correlation of its output with the actual number of sequences that folds onto the conformation being examined). Regression analysis is carried out using WEKA software.[40]

We construct a non-linear regression function based on the above features. The correlation between designability (i.e., the number of sequences that fold to a specific conformation) and the values returned by the nonlinear regression function is then calculated (see Table II). We construct regression functions using all of the features and taking each feature individually.

We observe a positive correlation between the topological arrangement of a conformation and its designability. Based on this result, we have then utilized these graph features to predict a range of designabilities instead of simply predicting a single designability value for a conformation. For this approach we provide a confidence interval for the predicted designabilities. The Naïve Bayes classifier is utilized for these predictions.

A. Outline of naïve Bayes prediction procedure

Given a hypothesis h and data D which bear on the hypothesis we have

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)},$$

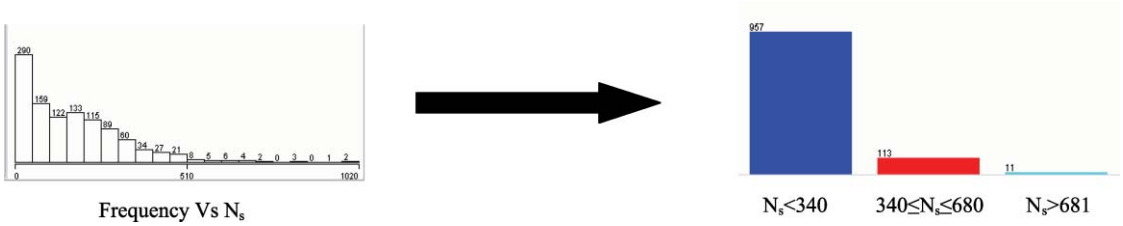

FIG. 2. Discretization of designability distribution into three bins using WEKA software. This corresponds to a simple accumulation of the individual peaks in the histogram on the left into three ranges given below the figure on the right.

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

The above relationship is the Bayes' theorem. A naïve Bayes classifier is a simple probabilistic classifier based on Bayes' theorem with *independence* assumptions. 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.

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. We first look at the distribution of designabilities for all the possible conformations for a particular model. We discretize the designabilities into three groups or bins such that the overall distribution of designability is preserved. This process of binning simplifies the calculations. We have also tried using higher numbers of bins and obtained comparable results.

This discretization is done using WEKA software. This procedure is demonstrated in Fig. 2 for an example of the 5×5 lattice.

In the training step, for each conformation described by 15 vectors or features, we calculate $P(\text{feature}_i|\text{range}_j)$, where $1 \leq i \leq 15$ and $1 \leq j \leq 3$ for the three selected bins. We also calculate $P(\text{range}_j)$ and $P(\text{feature}_i)$.

In the testing step, the $P(\text{range}_j|\text{feature}_i)$ is calculated using Bayes' theorem. Therefore, all of the features that define a conformation can be used together to predict the most probable range that its designability can fall to. A range for the designability value is predicted, and the prediction is considered "correct" if the actual designability value lies in that range. For each interaction graph we will calculate a confidence interval for its designability value.

If the actual designability of the conformation really falls within the predicted range of maximum probability, then the prediction is correct for that conformation. Therefore, the accuracy of a prediction can be calculated for each lattice model of interest by finding the correctly predicted instances. We utilize ten-fold cross validation to estimate the performance of the classification scheme while minimizing over-fitting (i.e., reducing the possibility of biased predictions).

III. RESULTS

We find the total number of different sequences that folds to each conformation having unique minimal (ground state) energy. Some conformations show high designability while the others are poorly designable. We have performed an analysis of H/P ratios of designable and poorly designable sequences for each lattice model (see supplementary Table S3).[39] The H/P ratios of designable sequences are always slightly larger than those of the poorly designable sequences. More H type residues imply stronger interactions with the present energies.

When regression analysis is conducted on the larger lattice models such as the $3 \times 3 \times 3$ and 5×5, we are able to get a correlation greater than 0.50 for the prediction using ten-fold cross validation. This is when all the topological features were employed. Here, a training set is used to obtain regression functions and these are then tested using other conformations to see if that regression function can predict the designabilities of these conformations. Results obtained for the lattice models are listed in Table II.

The correctly classified percentages when predicting the designability ranges using naïve Bayes classifier with all the features are shown in Table III. The overall prediction accuracy increases with the size of the lattice (see supplementary Fig. S2).[39] The accuracy of prediction is around 67% for the smallest 3×4 lattice model and it is the highest for the $3 \times 3 \times 3$ model, reaching almost 94%.

We search for a set of features that would give reasonable prediction of the designability range. By looking at the ranks of the importance of the features using correlation-based feature subset selection,[41] we find a set of important features for 2D lattices (*2D features*) and a similar set for 3D lattices (*3D features*). The selected 2D features were: *number of nodes with degree one, number of components, maximum shortest path length* and *number of nodes with degree equal to the average degree in the overall graph*. The set of 3D features selected were: *average shortest path, number of connected components,* and *number of nodes with maximum shortest path length*. We also searched for a *representative feature set* that would give a reasonable prediction of the designability range in both 2D and 3D lattices at the same time. These features

TABLE III. Prediction accuracy of designability for different lattice models.

Lattice	Prediction accuracy
3×4	67.7%
4×4	59.8%
5×5	80.9%
2×2×3	72.6%
3×3×3	93.8%

TABLE IV. AUCs for different feature sets for varying ranges of designabilities.

Lattice	Range	With all features	With 2D feature set	With 3D feature set	With representative set
3×4	$N_s < 25$	0.71	0.6	0.48	0.69
	$25 \leq N_s \leq 48$	0.89	0.86	0.58	0.87
	$N_s > 48$	0.7	0.82	0.63	0.74
4×4	$N_s < 188$	0.79	0.81	0.64	0.77
	$188 \leq N_s \leq 354$	0.66	0.66	0.62	0.64
	$N_s > 354$	0.80	0.81	0.72	0.69
5×5	$N_s < 340$	0.65	0.67	0.56	0.65
	$340 \leq N_s \leq 680$	0.65	0.66	0.53	0.65
	$N_s > 681$	0.81	0.75	0.7	0.86
2×2×3	$N_s < 15.7$	0.65	0.66	0.61	0.43
	$15.7 \leq N_s \leq 31$	0.63	0.63	0.6	0.4
	$N_s > 31$	0.75	0.77	0.65	0.56
3×3×3	$N_s < 394$	0.84	0.79	0.81	0.8
	$394 \leq N_s \leq 787$	0.84	0.78	0.81	0.8
	$N_s > 787$	0.87	0.87	0.82	0.85

included *average shortest path length, number of nodes with average degree, number of nodes with minimum shortest path length, number of nodes with average shortest path length, number of nodes with degree 1 and number of nodes with degree 2*.

We looked at the receiver operation characteristic (ROC) curve in which the true positive rate (sensitivity) is plotted as a function of the false positive rate (1-specificity) of predictions. The area under the ROC is given by AUC and is a measure of the prediction accuracy of the classifier. AUC can take values between 0 and 1; the closer this value is to 1, the better the prediction accuracy of the classifier. We looked at the AUCs for each model with each of the above sets of features. Results are shown in Table IV. Area under the curve is a standard way to assess the performance, and the higher the AUC value the better the prediction.

We are mostly interested in the *highly designable* range (range 3). We wanted to see whether the selected features can be used to recognize the highly designable conformations. For 3×4, 4×4, 5×5, $2 \times 2 \times 3$, and $3 \times 3 \times 3$ the AUCs for the designable range are 0.70, 0.80, 0.81, 0.75, and 0.87, respectively. As the size of the lattice increases so do the AUCs enabling the distinction of designable conformations (see supplementary Fig. S3).[39]

IV. DISCUSSION

We have shown that there exists a positive relationship between some graph features and designability of conformations. Furthermore, we have shown that the prediction may be improved by utilizing a confidence interval instead of relying upon a single value of designability. Moreover, we show that highly designable structures can be distinguished from other non-designable conformations based on the interaction graph features. This finding confirms the fact that the topology of a conformation is an important determinant in its designability.

We also found that the results improve as the size of the lattice increases. Very small lattice proteins exhibit anomalies due to very high fraction of surface residues, that (especially those located at corners in 2D) have very limited possibilities of forming non-bonded contacts. We observe that the size of the model used is more important than its dimensionality (see supplementary Figs. S1–S3).[39] It should be noted that this observation is based only on a small number of cases for small lattice proteins and therefore cannot be generalized.

For some cases in Table IV, the AUC for all the features is less than that with the selected set of features. For example, consider the AUC values for 3×4, $Ns > 48$ case. The AUC for all the features is 0.70 whereas for the 2D set of features it is 0.82. We believe that this is due to the violation of the independence assumption of the naïve Bayes classifier; some features are not independent of each other, and thus these features can cause the classifier to be biased in the favor of the redundant features.

We have carried out a correlation study on each pair of features used. A few features show a high correlation and most features are only slightly correlated with each other (see supplementary Table S1).[39]

So far we have used only the interaction graphs in our studies. However, there are many other graph representations that could be used to represent conformations. Use of line graphs is just another example. Here each vertex of the conformations is represented by an edge in the corresponding line graph and two vertices of the conformation are adjacent if and only if their corresponding edges share a common end point in the line graph. We expect to use line graphs to represent conformations and repeat what we did with the interaction graphs. Similar graph features can be obtained and checked to see if they are able to predict the conformational designabilities. Furthermore, different graph representations may be combined either by the classifier or using a graph tensor product to obtain new representation of the folds in lattice space. We hope to use other similar graph representations and also expect to employ product of graphs to come up with a suitable representation of conformations.

However, since designability is driven by minimizing protein energy that is computed from nearest neighbor non-bonded interactions, the "interaction graph" is more natural for the analysis than the "line graph." Nevertheless different representations of the same object are useful, even if they

carry similar information. However, some representations are more natural for the analysis of a given phenomenon than others.

A. Application of these features to find designable conformations of larger lattices and real structures

It is computationally infeasible to enumerate all the conformations of larger lattices. Thus far, the largest lattice space for which complete enumerations of Hamiltonian walks have been performed is for the $3 \times 4 \times 4$ lattice.[42] If a set of features can be found to predict the designability of a conformation, then that information can be used in generating conformations within a lattice space such that they are highly likely to be designable without having to generate the full conformation space.

In order to utilize the graph features to reduce the search-space for valid protein confirmations, random sampling of conformations of a smaller lattice for which complete enumeration can be done may be conducted to select a set of important features for that particular structure from the randomly sampled conformations. Such features may then be utilized to predict the highly designable conformations that have not been enumerated in random sampling. Since the complete enumeration of structures is possible for these smaller lattices, highly designable conformations may be compared to find out whether predicted designable conformations are indeed designable. Depending on the success of this method we can further do random sampling of conformations of larger lattice spaces and predict the possible designable conformations.

Going further, we can do a similar analysis for real protein structures (off-lattice models of proteins). This study will enable us to investigate the relationship between real protein topologies and their designabilities. We hope to carry out this analysis in the future.

ACKNOWLEDGMENTS

We would like to acknowledge support from NIH Grant Nos. R01GM072014, R01GM073095, R01GM081680, and R01GM081680-S1.

[1]F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, J. Mol. Biol. **112**(3), 535 (1977).
[2]C. Chothia, Nature **357**(6379), 543 (1992).
[3]Y. I. Wolf, N. V. Grishin, and E. V. Koonin, J. Mol. Biol. **299**(4), 897 (2000).
[4]K. Illergard, D. H. Ardell, and A. Elofison, Proteins: Struct., Funct., Bioinf. **77**(3), 499 (2009).
[5]P. Wong and D. Frishman, PLoS Comput. Biol. **2**(5), 392 (2006).
[6]H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**(5275), 666 (1996).
[7]H. Cejtin, J. Edler, A. Gottlieb, R. Helling, H. Li, J. Philbin, N. Wingreen, and C. Tang, J. Chem. Phys. **116**(1), 352 (2002).
[8]R. Helling, H. Li, R. Melin, J. Miller, N. Wingreen, C. Zeng, and C. Tang, J. Mol. Graphics Modell. **19**(1), 157 (2001).
[9]J. Y. Yang, Z. G. Yu, and V. Anh, J. Chem. Phys. **126**(19), 195101 (2007).
[10]R. Melin, H. Li, N. S. Wingreen, and C. Tang, J. Chem. Phys. **110**(2), 1252 (1999).
[11]C. Tang, Physica A **288**(1), 31 (2000).
[12]J. Miller, C. Zeng, N. S. Wingreen, and C. Tang, Proteins: Struct., Funct., Genet. **47**(4), 506 (2002).
[13]D. G. Covell and R. L. Jernigan, Biochemistry **29**, 3287 (1990).
[14]E. G. Emberly, J. Miller, C. Zeng, N. S. Wingreen, and C. Tang, Proteins: Struct., Funct., Genet. **47**(3), 295 (2002).
[15]A. R. Atilgan, P. Akan, and C. Baysal, Biophys. J. **86**(1), 85 (2004).
[16]G. Bagler and S. Sinha, Phys. A Stat. Mech. Appl. **346**(1–2), 27 (2005).
[17]N. V. Dokholyan, L. Li, F. Ding, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **99**(13), 8637 (2002).
[18]L. H. Greene and V. A. Higman, J. Mol. Biol. **334**(4), 781 (2003).
[19]A. Kloczkowski and R. L. Jernigan, Macromolecules **30**, 6691 (1997).
[20]K. V. Brinda and S. Vishveshwara, Biophys. J. **89**(6), 4159 (2005).
[21]K. S. Ramesh, K. V. Brinda, and S. Vishveshwara, Proteins: Struct., Funct., Bioinform. **59**(3), 616 (2005).
[22]A. N. Jha, G. K. Ananthasuresh, and S. Vishveshwara, PLoS ONE **4**, 8 (2009).
[23]S. Vishveshwara, K. V. Brinda, and N. Kannan, J. Theoret. Comput. Chem. **1**, 187 (2002).
[24]V. Shahrezaei and M. R. Ejtehadi, J. Chem. Phys. **113**, 6437 (2000).
[25]J. L. England, B. E. Shakhnovich, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **100**(15), 8727 (2003).
[26]T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. U.S.A. **101**(21), 7960 (2004).
[27]J. R. Banavar, T. X. Hoang, A. Maritan, F. Seno, and A. Trovato, Phys. Rev. **70**, 041905 (2004).
[28]A. Kloczkowski and R. L. Jernigan, Comput. Theoret. Polym. Sci. **7**, 163 (1997).
[29]A. Kloczkowski, T. Z. Sen, and R. L. Jernigan, Polymer **45**, 707 (2004).
[30]A. Kloczkowski and R. L. Jernigan, J. Chem. Phys. **109**, 5147 (1998).
[31]A. Kloczkowski and R. L. Jernigan, J. Chem. Phys. **109**, 5134 (1998).
[32]M. Peto, A. Kloczkowski, V. Honavar, and R. L. Jernigan, BMC Bioinf. **9**(1), 487 (2008).
[33]M. Peto, T. Z. Sen, R. L. Jernigan, and A. Kloczkowski, J. Chem. Phys. **127**, 044101 (2007).
[34]H. S. Chan and K. A. Dill, Phys. Today **46**(2), 24 (1993).
[35]D. J. Lipman and W. J. Wilbur, Proc. R. Soc. Lond. Ser. B: Biol. Sci. **245**(1312), 7 (1991).
[36]M. Peto, A. Kloczkowski, and R. L. Jernigan, J. Phys.: Condens. Matter **19**, 285220 (2007).
[37]K. A. Dill, Protein Sci. **8**, 1166 (1999).
[38]H. Li, C. Tang, and N. S. Wingreen, Proteins: Struct., Funct., Genet. **49**, 403 (2002).
[39]See supplementary material at http://dx.doi.org/10.1063/1.3596947 for Tables S1–S3 and Figs. S1–S3.
[40]M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, SIGKDD Explor. **11**(1), 10 (2009).
[41]M. A. Hall and L. A. Smith, "Practical feature subset selection for machine learning," in *Proceedings of the 21st Australasian Computer Science Conference* (Springer, New York, 1998), pp. 181–191.
[42]V. Pande, C. Joerg, A. Y. Grosberg, and T. J. Tanaka, Phys. A: Math. Gen. **27**(18), 6231 (1994).