Understanding Protein Structure from a Percolation Perspective

Dhruba Deb,[†] Saraswathi Vishveshwara,[†] and Smitha Vishveshwara[‡*]
[†]Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India; and [‡]Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois

ABSTRACT Underlying the unique structures and diverse functions of proteins are a vast range of amino-acid sequences and a highly limited number of folds taken up by the polypeptide backbone. By investigating the role of noncovalent connections at the backbone level and at the detailed side-chain level, we show that these unique structures emerge from interplay between random and selected features. Primarily, the protein structure network formed by these connections shows simple (bond) and higher order (clique) percolation behavior distinctly reminiscent of random network models. However, the clique percolation specific to the side-chain interaction network bears signatures unique to proteins characterized by a larger degree of connectivity than in random networks. These studies reflect some salient features of the manner in which amino acid sequences select the unique structure of proteins from the pool of a limited number of available folds.

INTRODUCTION

Anfinsen's landmark discovery (1) that the three-dimensional structure of protein is encoded in the amino acid sequence was made more than three decades ago. Although enormous progress has taken place in decoding the principles of protein folding, a definite scenario, as in the case of the identification of triplet genetic code for amino acid sequence in proteins (2–4) has not yet emerged. This is due to the fact that several factors such as the random and the selective behavior of the poly-peptide chain, optimization of geometry and energy play a role in the folding of proteins to their unique native state (5,6). Additionally, evolution has played a major role in selecting proteins, whose structures are optimized for functioning in their environment. Hence, the optimization of any specific parameter could have taken place to the extent of necessary and sufficient level and not necessarily to the maximum extent. Many important investigations have been carried out for several decades addressing different aspects. The selection of secondary structures due to geometric constraints (7), the geometry optimization model (5) and the energy landscape model (6) are a few examples. Furthermore, the availability of a large number of protein structures has aided in formulating and testing the proposed hypotheses. In this study, we have investigated the network of connections made by noncovalent interactions within the proteins, with a focus of identifying random as well as selective regimes in the network.

It is well known that proteins respect severe constraints imposed by folding entropy (7) and their backbone is arranged in regular arrays of secondary structures such as helices and sheets (8). The backbone endows the protein a robust skeletal structure composed of optimally packed, immutable folds (8–11) that are resilient to local variations and mutations (12,13). Furthermore, extensive sequence-structure correlation studies have shown a diversity of sequences for a given backbone structure. However, the underlying global structure of amino acid linkages formed via noncovalent side-chain interactions, which are also known to be crucial for the stability and uniqueness of protein structure, has received much less attention (14). The element of randomness at the noncovalent interaction level has been investigated at a preliminary level by considering the protein structures as networks (15) (K. V. Brinda, S. Vishveshwara, and S. Vishveshwara, unpublished data).

In this study, we have constructed structure networks (graphs) of several proteins based on the noncovalent interactions, both at the backbone level as well as including all the atoms of the side chains. The network parameters obtained from such graphs are compared with different random models, ranging from the most basic, unconstrained random model (Erdős-Rényi (ER)) to the ones constrained to mimic the protein topology. We specifically compare the percolation behavior of the protein with those of the random graphs by investigating the percolation of basic connections (bond percolation) (16) as well as higher order connections (clique percolation) (17). We find a striking resemblance between the bond percolation of the protein and all the random models. Additionally, we also find that the clique-percolation profile of the protein backbone connection graph resembles those of the random graphs. Interestingly, the protein side-chain connectivity graph exhibits clique percolation, which does not take place in any of the random models. Furthermore, we also observe such a percolating clique in decoy structures, which are poor in secondary structures and represent the molten globule state (18,19). By our study, we have been able to distinguish the side-chain connectivity in well packed secondary structures as the selective feature unique to folded proteins in their native state. Thus, the protein adopts the unique fold/structure in which the sequence is capable of making a percolating clique. In other words, the side chains interact in a highly connected

Submitted April 24, 2009, and accepted for publication July 15, 2009.

*Correspondence: smivish@illinois.edu

Editor: Ruth Nussinov.

0006-3495/09/09/1787/8 $2.00

doi: 10.1016/j.bpj.2009.07.016

fashion, stitching different secondary, super-secondary structures and stabilizing the protein structure at the global level. Our results are consistent with the fact that diverse sequences carrying out a variety of functions can adopt the same fold. We have considered the ubiquitous fold of TIM barrel (α/β fold), which is taken up by a large number of dissimilar sequences carrying out diverse functions, the Helix bundles (all-α) and the Lectins (all-β). We show that the commonality between them is a percolating clique of side-chain connectivity, which link different secondary and super-secondary structures.

METHODS

Data set

The data set used for this analysis on the general features consists of a set of 50 single-chain proteins (10 proteins for each size of 200, 400, 600, 800, and 1000 amino acids) with known structures obtained from the Protein Data Bank (20) (Table S4 in the Supporting Material). To investigate the fold specific features we have considered a data set of 15 proteins (five proteins for each of the folds: α/β, all-α and all-β) obtained from the Protein Data Bank (Table S5). The decoy structures were taken from Decoys 'R' Us database (18).

Networks and percolation theory

Much of the analysis of the protein network is based on key concepts borrowed from complex network theory and percolation studies. Broadly, a network (graph) consists of a collection of points (nodes) connected to one another by bonds (links). The nature of the network and the degree to which it is connected largely depends on the guiding principles governing the formation of links; for a class of random networks the formation of a link depends on a given probability of connection. The links, for instance, depend on the noncovalent connections in the case of protein structures and on the interacting proteins in protein-protein interaction network. A signature feature identifying properties of a network is the degree distribution, the degree being the number of links connected to a node. For example, a large class of random networks is known to exhibit degree distributions that peak around a specific value. On the other hand, some of the real-world networks such as the protein-protein interaction network or the spread of diseases (21,22), exhibit scale free networks or small-world network behavior in which certain nodes are highly connected.

The hallmark of a broad class of random networks is the presence of a transition point at which a giant connected cluster percolates the system whereas below this threshold (critical point), only smaller clusters are present. At the simplest level, the giant cluster may consist of connected bonds and the transition point can be identified by the size of the largest cluster as a function of the probability of connections. Instead of a simple bond percolation, we can envisage the percolation of more densely connected object-clique percolation. A clique, in a network, is a cluster where each node is connected to every other node. If the number of nodes in a clique is k, a community is defined as the collection of adjacent k-cliques where each clique shares k-1 nodes with the adjacent clique (17). Hence the largest community, which spans over the entire network, is a percolated clique and we use the terminology of ''largest community'' for clique percolation.

Representation of protein structures as networks

Protein side-chain network (PScN) is constructed on the basis of the details of the side-chain interactions, which is quantified in terms of the extent of interaction (23). Protein backbone network (PBN) is constructed by considering the Cα atom of each residue in the protein as a node and any two Cα atoms (excluding the sequence neighbors) situated at a distance less than a cut-off distance are connected by an edge (24). A brief description of this method is provided in the Supporting Material. The principle behind construction of PScN and PBN is pictorially depicted in Fig. 1. In this study, we identify the number of connections in PBN as a function of Cα-Cα distance ranging from 4.5 Å to 10 Å and I_{min} ranging from 1% to 9% in PScN.

Random network models

Three types of random graphs are used for comparison with the protein graphs. One of the models (RM1) is a simple unconstrained model similar to that of ER. The second one (RM2) is constrained to the topology of the protein, which obeys the rule of excluded volume. The third one (RM3) is the same as RM2, except that the node (amino acid) position is also constrained to that of the protein.

ER random network model (RM1) and mapping of connection to probability

The ER model is arguably the best studied model for random networks. It has the simple feature that any node can be linked to any other with some probability p. Several features of this model are known analytically. In particular, its degree distribution for a number of links k follows a Poisson curve $n(k) = N\,(pN)^k\,e^{-pN}/k!$, where N is the total number of nodes, and the critical probability for the bond percolation transition is at $p = 1/N$. For the k-clique percolation transition, critical probability is at $p(k) = 1/[(k-1)N]^{1/(k-1)}$. Based on compelling trends that we observed in protein structure, we have used the ER model and variants thereof to compare with the network properties of proteins.

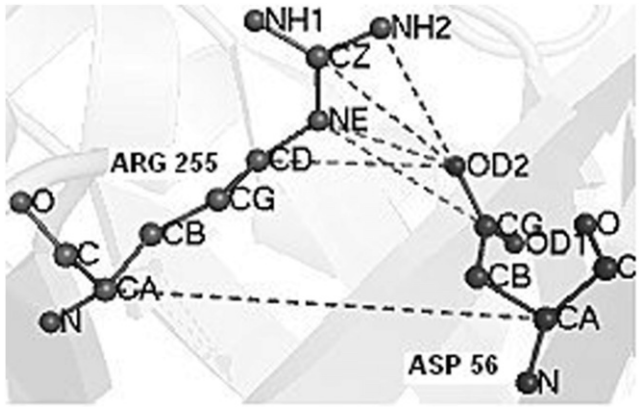

FIGURE 1 Representation of noncovalent connections for the protein backbone (PBN) and the side-chain (PScN) graphs. Two amino acids (ARG255 and ASP56) are shown in ball and stick model in the protein dihydropteroate synthase from *Escherichia coli* (Protein Data Bank (PDB) ID = 1AJ0). Cα atoms are separated by 8 Å (*dashed line*) and the two residues are considered as connected in PBN when Cα distance cutoff is <8 Å. Five pairs of atom-atom contacts among the side chains are well under 4.5 Å (3.38 Å, 3.5 Å, 2.78 Å, 3.74 Å, and 3.77 Å shown as *dashed lines*) and it corresponds to an interaction value (I_{ij}) of 6% according to the following equation: $I_{ij} = (n_{ij}/\sqrt{(N_i \times N_j)}) \times 100$, where, n_{ij} is the number of distinct atom pairs between the side chains of amino acid residues i and j, which are within a distance of 4.5 Å, and N_i and N_j are the normalization factors obtained from a nonredundant data set for residue types i and j. Any two residues, having interaction greater than a specified value (interaction cutoff, I_{cutoff} or I_{min}), are connected by an edge in the graph. Thus an edge in PScN is made if I_{min} is <6%.

We selected random graphs of the node sizes 200, 400, 600, 800, and 1000 to represent proteins of different sizes. Several realizations of ER random graphs were generated for the given node size with varying probability of edges. The number of edges (the average of 10 proteins of chosen size and obtained under a given condition) in the protein graph is matched to the corresponding probability of connection in the ER graph. Thus, the number of edges is matched to the probability of connection.

Constrained random network models

Finite size random node-constrained random edge model (RM2)

Proteins are of finite size and the RM1 model, which is not constrained in space, is not the best random model to compare the protein structure networks. Hence we have constructed random models, which are constrained to finite size, idealized to spherical shape to mimic the shape approximately taken up by globular proteins. In this model, the nodes are generated randomly within a sphere, the radius of which is chosen as approximately the average radius of gyration (R) from the data set of globular proteins of selected size. Hence each of the node coordinate (x,y,z) is within the spherical limit of R. The random model thus constructed, exhibits a compactness similar to real proteins, as the radius of gyration is a measure of compactness of protein (25). The specified numbers of edges (corresponding to the number found in protein of the selected size in both the PBN and PScN) are distributed randomly among a pair of nodes, which are within a distance of 6.5 Å or 7.5 Å, or 8.5 Å in three-dimensional space. A distance of 6.5 Å corresponds to the first peak in the radial distribution of residues in the interior of proteins (26,27). However, 7.5 Å, or 8.5 Å distances are also used not to ignore any atom-atom contact (see Fig. S3). Second, stearic contact is avoided by not connecting the nodes, which are within 4.5 Å of each other. Such a model is protein-like in its size, has realistic connections in space, and respects the excluded volume criterion. This model is averaged over 20 random realizations.

Protein nodes constrained random edge model (RM3)

The RM2 model mentioned above captures many features of proteins and is a generalized model applicable to a large number of globular proteins. However, it deviates from the exact size and does not follow the chain connectivity. These features can be incorporated in a protein specific model, by keeping the nodes of the random graph identical to that of the selected protein and randomly rewiring only the edges. To make realistic edges, the specified number of connections (corresponding to the number found in protein of the selected size in both the PBN and PScN) are randomly distributed within a physical distance (4.5 Å $<$ distance $<$ 6.5 Å or 7.5 Å or 8.5 Å) of each amino acid in the protein structure. Because the number of edges within a sphere of 6.5 Å is much greater than the maximum number found in the PScN for a given node size (see Table S2), it is possible to randomly distribute the edges of smaller number. In the case of PBN, the number of edges corresponding to a lower cutoff (4–9 Å) is selected randomly from the repertoire of edges obtained from a cutoff of 10 Å. In this way, 10 realizations for each protein in the data set are created and finally evaluated parameters are averaged over each of the 10 proteins in the data set. We denote this model as RM3 model. If proteins are optimally packed with secondary and super-secondary structures, irrespective of the side chain (5), this model provides a reference point to test the exclusive role played by side-chain interaction because the topology of the model is strictly constrained to that of the protein.

Community identification

For community identification, we have used the program CFinder (v.1.21) (28). An example of k-clique ($k = 3$) community in the PScN (protein dihydropteroate synthase from *Escherichia coli* at $I_{min} = 3\%$) is shown in Fig. 2.

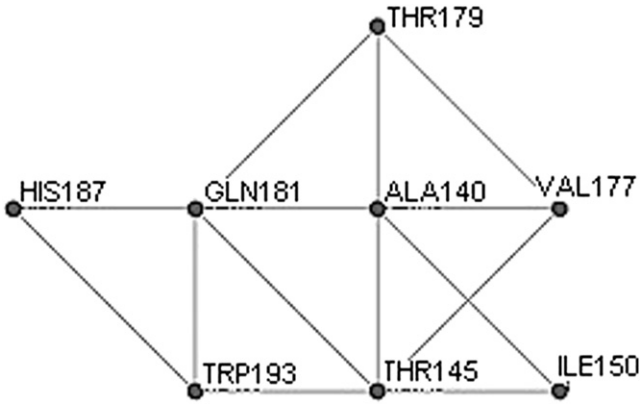

FIGURE 2 Largest k-clique ($k = 3$) community in the dihydropteroate synthase from *Escherichia coli* (PDB ID = 1AJ0) at $I_{min} = 3\%$.

RESULTS

Protein structure and the random networks

Two types of protein structure graphs have been investigated in this study. The PBN represents the polypeptide chain packing and the PScN focuses on the details of side-chain interactions in the proteins. From the network point of view, the number of connections for a given node size differs depending on the criteria used for connections. For example, proteins of the size of ~400 amino acids make 396–3679 number of Cα-Cα connections, when the residues with in a range of 4.5–10 Å are considered to be connected in PBN. Similarly, the number of connections for a 400 residue protein varies from 798 to 133 in PScN, depending on the side-chain connection strengths ranging from I_{min} of 1–9% (see Table S2). An important difference to notice between PBN and PScN is that the PBNs accommodate more number of edges than the PScNs. There is very little overlap between the number of connections of backbone and the side-chain regimes. The number of edges plays a significant role in the corresponding random graphs because the likelihood of percolation increases with an increase in the probability of connections and one can comfortably separate the random graphs as PBN or PScN like. For the sake of brevity, we have presented the results pertaining to the node size of 400, although qualitatively the same results are obtained for other sizes. (Some important results for other sizes are presented in the Supporting Material.) We characterize the PBN and PScN in terms of their degree distribution and compare them with the three random models. Next, we examine the percolation behavior at the simple bond-connection level and then at the clique-connection level.

Degree distribution

It is noteworthy that the degree distribution of PBN and PScN follow approximately the same behavior as that of the RM1 model at different levels of connections (see

Fig. S1). The degree distribution plots of PScN fit best to the Poisson distribution (see Fig. S2) and this rules out scale-free behavior in protein structure networks. They do differ slightly from RM1 model. For example, the Poisson fitting parameters are different for RM1 and PScN (see Table S1). Additionally, the number of orphan nodes, which are not connected to any other node in the network, is higher in protein structure network than RM1 (see Fig. S1). The RM2 and RM3 models as expected exhibit the degree distribution behavior closer to the protein case, with increased number of orphan nodes compared to RM1. Thus, there is an element of randomness in the noncovalent interactions within proteins. However, a larger number of orphan nodes in the protein case imply more connections in the connected regions, as the total number of nodes and edges are comparable for the protein and the random graphs. Although this effect does not cause any drastic change at the degree distribution level, the effect of this can be seen in clique percolation, as discussed in a later section.

Bond percolation

In this study, we characterize the percolation properties of proteins based on our reference random networks. We compare the sizes of the largest clusters in protein structure networks to those of the reference networks as a function of probability of edge formation.

As mentioned earlier, the key factor is the number of edges that a protein can make, depending on the definition of contact. There is an inherent limitation to connections in proteins, due to factors like excluded volume, the nodes being connected as a polymer chain, and the geometry adopted by proteins. We adhere to the number of connections in protein graphs while constructing the random graphs. (However, the number of connections is expressed as the probability of connection as given for 400 node graphs in Table S2.) The only freedom we exercise is to distribute the nodes and the edges randomly or in a constrained manner as described in the Methods section.

Bond percolation behavior is examined by plotting the size of the largest cluster as a function of the probability of connection. In the PBN, the size of the largest cluster reaches a maximum (size of the number of nodes) at a probability of connection being 0.006 (corresponding to Cα distance cutoff of 5 Å) as shown in Fig. 3 *b*. Even at the minimum possible probability of connection (Cα distance cutoff of 4.5 Å), the size of the largest cluster is very close to that of the maximum. This implies that the percolation at the backbone level is almost complete at the minimum realistic probability of connection. Strikingly, the size of the largest cluster is obtained at around the same probability of connection in RM2 and RM3, indicating that the backbone connections in a random model obeying the constraints of protein topology and excluded volume exhibits the features of the protein graph. The size of the largest cluster in RM1, however, reaches the maximum at an increased probability of connection of 0.02 and the percolation transition also starts at a higher probability of connection than that of the protein and in the random models RM2 and RM3. The side-chain graph (PScN) on the other hand can take up much less number of connection, Here the maximum size of the largest cluster is slightly smaller than that of the node size, due to the existence of orphan nodes at all levels of probability (I_{min}) (Fig. 3 *a*). This is achieved around a probability of 0.01 ($I_{min} = 1\%$) and the bond percolation transition takes place around the probability of 0.005 ($I_{min} \sim 4\%$). As expected, the behavior of the constrained random models RM2 and RM3 is very close to that of the protein. The onset of percolation transition and the attainment of the largest cluster on the other hand are shifted to higher probabilities connections in RM1. Thus, the proteins behave random-like in their bond percolation feature, which is quite evident by almost identical behavior of random models constrained to protein geometry.

Clique percolation

In recent years, clique percolation transition is being used to uniquely identify local structural units of the real-world networks where more densely connected regions are considered to be essential in making predictions about yet unknown functions of proteins (28). Here too, such a percolation study serves to pinpoint the denser connectivity of the largest cluster of the protein structure network. We observe the

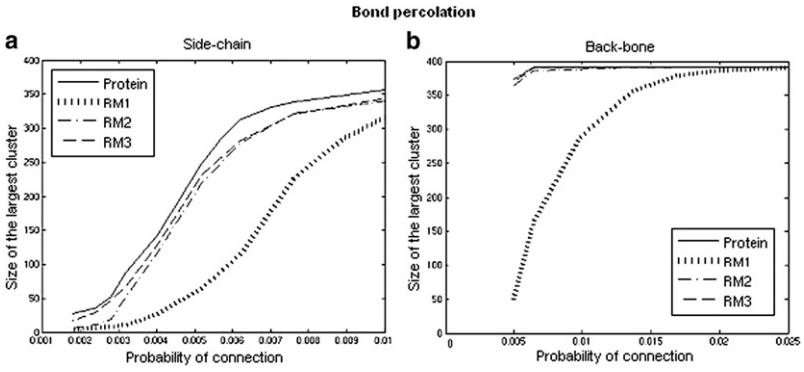

FIGURE 3 Largest cluster profile (averaged over 10 realizations for size 400 nodes) of (*a*) PScN and corresponding random models: RM1, RM2, and RM3, (*b*) PBN and corresponding random models: RM1, RM2, and RM3. In the side-chain profile, both PScN and RM1 show transition where sharp increase in the size of the largest cluster gives the curve its sigmoidal nature. However, RM1 has bond percolation at a higher probability of connection than PScN. In the backbone profile, at a high probability of connection, PBN has already reached the saturation for size of the largest cluster. But, RM1 shows partial transition on set of saturation. RM2 and RM3 show values in between protein and RM1 in side-chain profile whereas both the constrained models merge with protein in the backbone profile.

behavior of the largest community of k-clique as a function of probability where $k = 3$ (Although we obtain cliques of larger sizes, a large percolating community is obtained only for $k = 3$ in proteins. Therefore, all the clique percolation studies are carried out at $k = 3$ and the largest community is defined only for this case). In the backbone profile (Fig. 4 *b*), the probability range captures complete clique percolation transition of PBN and partial transition for RM1. Obviously, an uncorrelated random network requires more number of edges to attain a saturated community, which falls out of the backbone probability range. The RM2 and RM3 models with the protein geometry and topology constraints move closer to the PBN than the RM1 model as anticipated. The side-chain profile (Fig. 4 *a*), however, quite strikingly distinguishes PScN from all other reference networks. At a probability of 0.01 ($I_{min} \sim 1\%$), the largest community for PScN shows beginning of percolation transition with a steep increase in the community size. (It is to be noted that the community size in PScN will not reach the maximum of node size as in the case of PBN, even at the maximum possible probability of side-chain connections in proteins.) In contrast, the RM1 and even the constrained models RM2 and RM3 do not start percolating at all even at the maximum possible connection (atom-atom connection) level. An increase in the constraint by significantly decreasing outer topological boundary of nodes (from 8.5 Å to 6.5 Å, which effectively reduces the random selection of edges) also does not result in the onset of clique percolation. The result discussed here for the 400 node size is a general phenomenon common to proteins of all sizes. Relevant results for 200 and 600 node sizes are presented in Fig. S5.

The decoy structures simulated from the native structures have been generally associated with the molten globule state (18,19). We have examined the side-chain percolating communities in a set of 10 decoys for each of the 10 proteins (see Table S6). We observe that they have features common to those of native structures and they differ mainly by their reduction in the secondary structural content. The relevance of this result is discussed in the Discussion.

Clique percolation in proteins of different folds

The fact that amino acid sequence dictates the structure of proteins is well accepted in molecular and structural biology. The structures of >50,000 proteins have been resolved (20) and it has been possible to model the structures of new sequences using the available structures as templates (29,30). The success rate of modeling is high when there is high sequence similarity (>30%) with proteins of known structure. There are many structures (folds), however, that are taken up by a large number of sequences with a similarity as low as one can get by chance. The conventional methods of modeling fail in such a situation because there is no unifying principle. From this study, we believe that the possibility of a percolating clique can be a common phenomenon to stabilize a given fold adopted by diverse sequences. Hence, in this section, we have elucidated the details of the percolating cliques, which stabilize all-α, all-β, and one of the widely adopted α/β folds (TIM barrel is adopted by a large number of protein sequences with low similarity). This observation also provides a rationale for the fact that a vast range of amino acid sequences take up a highly limited number of folds.

The α/β barrel fold, or known more commonly as TIM barrel fold, first discovered in the structure of the protein triose phosphate isomerase, is one of the most ubiquitous folds in nature and has been extensively studied for the understanding it provides of protein structure, function and folding (31–35). We observe two or more large percolating cliques (at $I_{min} = 3\%$) for the proteins of the TIM fold (Fig. 5 and Fig. S4). These communities become further connected when the probability of connection increases at the maximum possible side-chain connection ($I_{min} = 1\%$) (see Table S5). The resulting giant community spans over the whole protein connecting several secondary structural elements. We notice the diversity of residues taking part in the clique formation in different proteins of the same TIM barrel fold. Consequently, the overall size of the community is similar in each of the TIM barrel proteins though it differs significantly in its residue arrangements. Furthermore, the

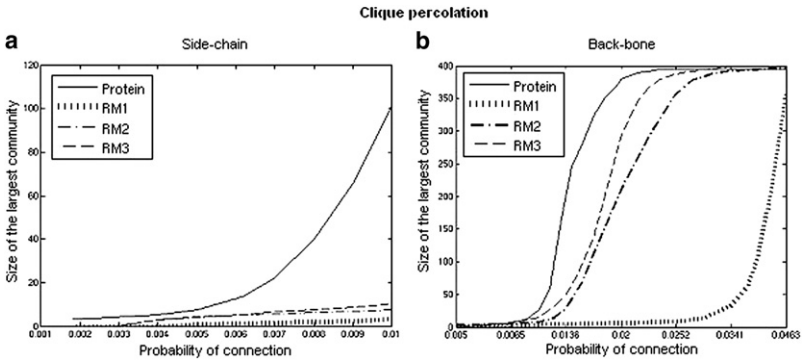

FIGURE 4 Clique percolation profile (averaged over 10 realizations for size 400 nodes) of (*a*) PScN, corresponding RM1 and constrained random networks, and (*b*) PBN, corresponding RM1 and constrained random networks. Number of nodes in the largest community is plotted as a function of probability of connection. In the community, each clique of size k nodes shares k-1 nodes with its adjacent clique. In this figure, cliques with value $k = 3$ are considered. The side-chain profile captures early stage of transition for PScN. However, RM1 has not entered the transition region in this probability range. On the other hand, the backbone profile, having a higher probability range, captures complete transition for PBN where size of the largest community shows a sharp increase giving the curve a sigmoidal nature. But, even this probability range is not enough to capture complete transition for RM1. The Rm2 and RM3 model in backbone profile behave almost similarly as protein network with the percolation transition at a little higher probability. But, at side-chain profile both the constrained models behave more similar to the random network.

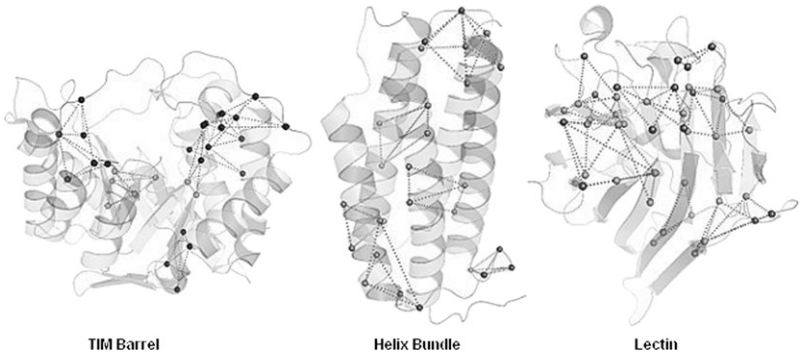

FIGURE 5 Clique percolation in TIM barrel protein dihydropteroate synthase (PDB ID = 1AJ0) (*left*), helix bundle protein cobalamin adenosyltransferase (PDB ID = 1NOG) (*center*), and lectin protein manganese concanavalin A (PDB ID = 1DQ6) (*right*) at $I_{min} = 3\%$. The helices, sheets, and loops are shown in cartoon representation. The residues involved in the formation of percolating clique are shown as spheres and the connections among them are shown as dashed lines. Only three residue cliques are shown here. The noncovalent connections among side chains of residues are shown as dotted lines. This figure shows that certain communities of varying sizes connect various secondary structures for all the three proteins at $I_{min} = 3\%$.

location of the percolating cliques in different proteins is different with respect to the overall geometry. Thus, the only feature common in all the TIM barrel folds is the occurrence of percolating side-chain cliques that stitch different secondary and super-secondary structures.

The helix bundle fold consists of several parallel or antiparallel α helices. In our study, we notice, unlike TIM barrel fold, five or six small percolating cliques (at $I_{min} = 3\%$) in all the proteins of helix bundle fold. With the increase in the probability of connection (at $I_{min} = 1\%$), these small communities get connected to each other resulting in a giant community (Fig. 5 and Table S5). In accordance with the results for TIM barrel fold, the giant community in helix bundle proteins spans over the whole structure linking the secondary structural elements.

The third fold we have studied is lectin, a well-known example of all-β fold. The communities observed in lectins (at $I_{min} = 3\%$) have varying sizes. We observe two or three large communities and several small communities (Fig. 5). As in the case of other two folds mentioned above, these communities connect each other to give rise to a giant community at the maximum possible side-chain connection ($I_{min} = 1\%$), which in turn, spans over the whole protein stitching the secondary structural elements.

In all the three folds, we observe diversity in residue type and arrangement involved in the formation of percolating cliques (see Table S5). The diversity in sequences is reflected in the composition and the architecture of these percolating cliques, thus accounting for the same fold adopted by dissimilar sequences and providing a rationale for limited conformational space.

DISCUSSION

The natural tendency of the polypeptide chain for the formation of secondary structures and their optimal packing limits the number of protein folds (5). On the other hand, the amino acid sequence in a protein uniquely determines the structure and hence the side chains and the order of their appearance in the chain ought to play a crucial role in selecting the unique structure. In other words, the folded structure of proteins is a result of the combination of certain statistically probable events and some selective events. In this study, we have addressed this issue by comparing the protein structure networks (made by the noncovalent connection both at the backbone level (PBN) and at the level including the details of the side chain (PScN)) with random models with and without realistic constraints such as protein topology and excluded volume. A simple bond percolation and an intricate connection of clique percolation are studied. The bond percolation at all levels of protein structure network resembles that of random networks. The clique percolation at the backbone level also resembles those of random models. On the other hand, only the protein side-chain network at the high level of connections (low I_{min}) is capable of clique percolation and none of the random models (including the one very similar to that of proteins) exhibited clique percolation. In general, clique percolation can take place in any system, given a large number of connections (17). The special feature of proteins is the existence of a percolating clique with a limited number of realistically possible connections, specifically atom-atom contact of noncovalently interacting side chains.

Optimal packing of secondary structures is also required for the uniqueness of proteins and it has been argued (36) that the polypeptide backbone inherently posses this feature. The percolating cliques of side chains, in addition to the packed secondary structures due to the backbone, confer uniqueness to the protein structure. An important issue with this regard is the manner in which molten globule structures differ from those of the native structures (37–42). The loss of secondary structures and a slight increase in the radius of gyration are considered to be the properties of molten globules. Computationally, decoy structures generated from the native structures have been considered to be equivalent of molten globule state. In this study, we have considered 10 decoy structures (18) of each of 10 different proteins (see Table S6) and compared the size of the largest community with those of the native structures. In most cases there is not much significant difference in terms of the size and there are substantial overlaps between the residues in the largest community of the decoys and their native states. (This may be due to the fact that the decoy structures are still in the conformational space close to that of the native.) However,

the percentage of secondary structures in the decoys has reduced significantly. Thus, it is clear that the uniqueness of the native states is due to both the optimal packing of secondary structures and their intactness preserved by a percolating community made up of the interactions of side chains.

Correlating the structure of proteins to their functions is an important goal of structural biologists. Experimentally, this aspect is probed by obtaining different complex structures of a given protein from x-ray crystallography and the dynamical structures are captured by NMR spectroscopy. Computationally, molecular dynamics simulations provide information by spanning the equilibrium conformational space. Because it is computationally expensive to carry out long time simulations, normal mode analysis (43–48) and elastic network models (ENM) (49–51) have been developed to extract meaningful dynamical modes from the static x-ray structures. ENM uses simplified potentials in which the C_α atom represents the residue, making the investigations of large system computationally accessible. ENM, which considers both the sequential and the special neighbors of a chosen residue in the polypeptide chain in their formalism, has done exceedingly well in characterizing complicated systems (52–57) due to the simplicity of the potential it uses. From this study, it seems that there is an important role played by the collective interaction of side-chain atoms. Further analyses would need to investigate whether the incorporation of an additional term in the ENM potential to represent the collective interactions of side chains in a simplified manner would further push ENM toward enhancing the accuracy of the model. Similarly, the concept of side-chain clique percolation can be incorporated in protein structure prediction methods to see if it improves the accuracy and/or the efficiency of the prediction.

In summary, it seems that the uniqueness of the protein structure is brought out by extremely specific side-chain interactions, along with well packed secondary structures. Our results are consistent with the sequence based statistical coupling analysis on evolutionary data on proteins (58,59). The nonbonded connections between side-chain atoms pervade the protein structure and stitch the secondary and super-secondary structures, stabilizing the fold taken up by the packing of the polypeptide chain. We have shown this feature in proteins belonging to three different folds. Thus, the key to the unique structure is indeed in the amino acid sequence, whereas the polypeptide backbone has given myriad structures to choose from. Although the protein sequence has the information to the protein fold in the form of percolating cliques of side-chain interactions, many sequences can hold the key to the same fold as shown in the case of diverse sequences belonging to the ubiquitous TIM barrel fold. Specifically, different combinations of the amino acid type and its position in the sequence, which can interact at the atomic level in a correlated fashion, are likely to stabilize the unique structure. This also provides a rationale for the fact that a vast range of amino acid sequences take up a highly limited number of folds.

SUPPORTING MATERIAL

Methods, five figures, and six tables are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(09)01238-7.

This work was supported by the National Science Foundation (DMR 06-44022 CAR) and Mathematical Biology project (DSTO773) funded by the Department of Science and Technology, India, for computational facilities.

REFERENCES

1. Anfinsen, C. B. 1973. Principles that govern the folding of protein chains. *Science*. 181:223–230.
2. Watson, J. D., and F. H. Crick. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*. 171:737–738.
3. Crick, F. H., L. Barnett, S. Brenner, and R. J. Watts-Tobin. 1961. General nature of the genetic code for proteins. *Nature*. 192:1227–1232.
4. Khorana, H. G., H. Buchi, H. Ghosh, N. Gupta, T. M. Jacob, et al. 1966. Polynucleotide synthesis and the genetic code. *Cold Spring Harb. Symp. Quant. Biol*. 31:39–49.
5. Trinh, X. H., A. Trovato, F. Seno, J. R. Banavar, and A. Maritan. 2005. Geometrical model for the native-state folds of proteins. *Biophys. Chem*. 115:289–294.
6. Miyashita, O., P. G. Wolynes, and J. N. Onuchic. 2005. Simple energy landscape model for the kinetics of functional transitions in proteins. *J. Phys. Chem. B*. 109:1959–1969.
7. Ramachandran, G. N., C. Ramakrishnan, and V. Sasisekharan. 1963. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol*. 7:95–99.
8. Przytycka, T., R. Aurora, and G. D. Rose. 1999. A protein taxonomy based on secondary structure. *Nat. Struct. Biol*. 6:672–682.
9. Chothia, C. 1992. Proteins. One thousand families for the molecular biologist. *Nature*. 357:543–544.
10. Denton, M., and C. Marshall. 2001. Protein folds: laws of form revisited. *Nature*. 410:417.
11. Hoang, T. X., A. Trovato, F. Seno, J. R. Banavar, and A. Maritan. 2004. Geometry and symmetry presculpt the free-energy landscape of proteins. *Proc. Natl. Acad. Sci. USA*. 101:7960–7964.
12. Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature*. 217:624–626.
13. King, J. L., and T. H. Jukes. 1969. Non-Darwinian evolution. *Science*. 164:788–798.
14. Greene, L. H., and V. A. Higman. 2003. Uncovering network systems within protein structures. *J. Mol. Biol*. 334:781–791.
15. Brinda, K. V., and S. Vishveshwara. 2005. A network representation of protein structures: implications for protein stability. *Biophys. J*. 89:4159–4170.
16. Stauffer, D. 1985. Introduction to Percolation Theory. Taylor and Francis, London, UK.
17. Derényi, I., G. Palla, and T. Vicsek. 2005. Clique percolation in random networks. *Phys. Rev. Lett*. 94:1–4.
18. Samudrala, R., and M. Levitt. 2000. Decoys 'R' Us: a database of incorrect protein conformations to improve protein structure prediction. *Protein Sci*. 9:1399–1401.
19. Yang, J. S., W. W. Chen, J. Skolnick, and E. I. Shakhnovich. 2007. All-atom *ab initio* folding of a diverse set of proteins. *Structure*. 15:53–63.
20. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, et al. 2000. The Protein Data Bank. *Nucleic Acids Res*. 28:235–242.

21. Albert, R., and A.-L. Barabasi. 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74:47–97.
22. Amaral, L. A., A. Scala, M. Barthelemy, and H. E. Stanley. 2000. Classes of small-world networks. *Proc. Natl. Acad. Sci. USA.* 97:11149–11152.
23. Kannan, N., and S. Vishveshwara. 1999. Identification of side-chain clusters in protein structures by a graph spectral method. *J. Mol. Biol.* 292:441–464.
24. Patra, S. M., and S. Vishveshwara. 2000. Backbone cluster identification in proteins by a graph theoretical method. *Biophys. Chem.* 84:13–25.
25. Sistla, R. K., K. V. Brinda, and S. Vishveshwara. 2005. Identification of domains and domain interface residues in multidomain proteins from graph spectral method. *Proteins.* 59:616–626.
26. Miyazawa, S., and R. L. Jernigan. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules.* 18:534–552.
27. Miyazawa, S., and R. L. Jernigan. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256:623–644.
28. Adamcsek, B., G. Palla, I. Frakas, I. Derényi, and T. Vicsek. 2006. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics.* 22:1021–1023.
29. Sali, A., and T. L. Blundell. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234:779–815.
30. Rohl, C. A., and D. Baker. 2002. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J. Am. Chem. Soc.* 124:2723–2729.
31. Kannan, N., S. Selvaraj, M. M. Gromiha, and S. Vishveshwara. 2001. Clusters in alpha/beta barrel proteins: implications for protein structure, function, and folding: a graph theoretical approach. *Proteins.* 43:103–112.
32. Lesk, A. M., C. I. Branden, and C. Chothia. 1989. Structural principles of alpha/beta barrel proteins: the packing of the interior of the sheet. *Proteins.* 5:139–148.
33. Wodak, S. J., I. Lasters, F. Pio, and M. Claessens. 1990. Basic design features of the parallel alpha beta barrel, a ubiquitous protein-folding motif. *Biochem. Soc. Symp.* 57:99–121.
34. Murzin, A. G., A. M. Lesk, and C. Chothia. 1994. Principles determining the structure of beta-sheet barrels in proteins. II. The observed structures. *J. Mol. Biol.* 236:1382–1400.
35. Bharat, T. A., S. Eisenbeis, K. Zeth, and B. Hocker. 2008. A beta alpha-barrel built by the combination of fragments from different folds. *Proc. Natl. Acad. Sci. USA.* 105:9942–9947.
36. Banavar, J. R., and A. Maritan. 2007. Physics of proteins. *Annu. Rev. Biophys. Biomol. Struct.* 36:261–280.
37. Kuwajima, K., H. Yamaya, S. Miwa, S. Sugai, and T. Nagamura. 1987. Rapid formation of secondary structure framework in protein folding studied by stopped-flow circular dichroism. *FEBS Lett.* 221:115–118.
38. Goto, Y., and A. L. Fink. 1990. Phase diagram for acidic conformational states of apomyoglobin. *J. Mol. Biol.* 214:803–805.
39. Chyan, C. L., C. Wormald, C. M. Dobson, P. A. Evans, and J. Baum. 1993. Structure and stability of the molten globule state of guinea-pig alpha-lactalbumin: a hydrogen exchange study. *Biochemistry.* 32: 5681–5691.
40. Luthey-Schulten, Z., B. E. Ramirez, and P. G. Wolynes. 1995. Helix-coil, liquid crystal and spin glass transitions of a collapsed heteropolymer. *J. Phys. Chem.* 99:2177–2185.
41. Levitt, M., M. Gerstein, E. Huang, S. Subbiah, and J. Tsai. 1997. Protein folding: the endgame. *Annu. Rev. Biochem.* 66:549–579.
42. Ferreiro, D. U., J. A. Hegler, E. A. Komives, and P. G. Wolynes. 2007. Localizing frustration in native proteins and protein assemblies. *Proc. Natl. Acad. Sci. USA.* 104:19819–19824.
43. Goldstein, H. 1950. Classical Mechanics. Addison-Wesley, Reading, MA.
44. Go, N., T. Noguti, and T. Nishikawa. 1983. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA.* 80:3696–3700.
45. Brooks, B., and M. Karplus. 1985. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proc. Natl. Acad. Sci. USA.* 82:4995–4999.
46. Levitt, M., C. Sander, and P. S. Stern. 1985. Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.* 181:423–447.
47. Ma, J. 2005. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure.* 13:373–380.
48. Van Wynsberghe, A. W., and Q. Cui. 2006. Interpreting correlated motions using normal mode analysis. *Structure.* 14:1647–1653.
49. Tirion, M. M. 1996. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* 77:1905–1908.
50. Bahar, I., A. R. Atilgan, and B. Erman. 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* 2:173–181.
51. Yang, L.-W., and C.-P. Chng. 2008. Coarse-grained models reveal functional dynamics - I. Elastic network models—theories, comparisons and perspectives. *Bioinform. Bio. Insights.* 2:25–45.
52. Hinsen, K. 1998. Analysis of domain motions by approximate normal mode calculations. *Proteins.* 33:417–429.
53. Bahar, I., B. Erman, R. L. Jernigan, A. R. Atilgan, and D. G. Covell. 1999. Collective motions in HIV-1 reverse transcriptase: examination of flexibility and enzyme function. *J. Mol. Biol.* 285:1023–1037.
54. Keskin, O., I. Bahar, D. Flatow, D. G. Covell, and R. L. Jernigan. 2002. Molecular mechanisms of chaperonin GroEL-GroES function. *Biochemistry.* 41:491–501.
55. Xu, C., D. Tobi, and I. Bahar. 2003. Allosteric changes in protein structure computed by a simple mechanical model: hemoglobin T<-->R2 transition. *J. Mol. Biol.* 333:153–168.
56. Cui, Q., G. Li, J. Ma, and M. Karplus. 2004. A normal mode analysis of structural plasticity in the biomolecular motor F(1)-ATPase. *J. Mol. Biol.* 340:345–372.
57. Wang, Y., A. J. Rader, I. Bahar, and R. L. Jernigan. 2004. Global ribosome motions revealed with elastic network model. *J. Struct. Biol.* 147:302–314.
58. Lockless, S. W., and R. Ranganathan. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science.* 286:295–299.
59. Dima, R. I., and D. Thirumalai. 2006. Determination of network of residues that regulate allostery in protein families using sequence analysis. *Protein Sci.* 15:258–268.