Graph theoretic properties of networks formed by the Delaunay tessellation of protein structures

Todd J. Taylor and Iosif I. Vaisman
Laboratory for Structural Bioinformatics, School of Computational Sciences, George Mason University, 10900 University Boulevard MSN5B3, Manassas, Virginia 20110, USA
(Received 31 October 2005; published 25 April 2006)

The Delaunay tessellation of several sets of real and simplified model protein structures has been used to explore graph theoretic properties of residue contact networks. The system of contacts defined by residues joined by edges in the Delaunay simplices can be thought of as a graph or network and analyzed using techniques from elementary graph theory and the theory of complex networks. Such analysis indicates that protein contact networks have small world character, but technically are not small world networks. This approach also indicates that networks formed by native structures and by most misfolded decoys can be differentiated by their respective graph properties. The characteristic features of residue contact networks can be used for the detection of structural elements in proteins, such as the ubiquitous closed loops consisting of 22–32 consecutive residues, where terminal residues are Delaunay neighbors.

DOI: 10.1103/PhysRevE.73.041925 PACS number(s): 87.15.By, 87.14.Ee, 36.40.Mr

I. INTRODUCTION

Many problems in protein structure analysis can be addressed using the representation of a protein structure as a residue contact map (for example [1–3]). In recent years a number of works have focused on the study of the topology and biological significance of networks formed by residues in contact [4–9]. In most of these works the contacts between residues are defined based on a pairwise residue separation in three-dimensional space, frequently relying on a somewhat arbitrary value of distance cutoff. In this paper we describe residue contact networks defined in a more robust way with Delaunay tessellation. We characterize the graph topology of these networks and show its applicability for identifying specific structural elements in protein architecture and its limited ability to discriminate between native and misfolded protein structures.

A. Delaunay tessellation of protein structures

A method of partitioning the space between a set of points known as Delaunay tessellation is gaining popularity in various protein structure analysis applications [7,10,11]. The analysis can be summarized in the following way. The protein is abstracted to a set of points. A point can correspond to an atom, a collection of atoms, or an entire residue. In the single point per residue representation, the coordinates of the point can be those of the α carbon, β carbon, or the center of mass of the side chain. In this work we use a single point per residue representation, where the points are located at α-carbon atoms. These points are joined by edges (Fig. 1) in a unique way to form a set of nonoverlapping, irregular, space-filling tetrahedra defined by the Delaunay simplices [12]. The tetrahedra have the property that the sphere on the surface of which all four vertices reside does not contain a vertex from any other tetrahedron (the empty sphere property). Residues joined by a Delaunay simplex edge are natural nearest neighbors in a well-defined sense [12]. The analysis of statistical characteristics of the tessellated protein structures has been used in fold recognition [13–15], for structure alignment and comparison [16–18], as a way to identify cavities in the surface of a protein that could be potential binding pockets [19], to study the stability and activity effects of point mutations [20,21], to define structural motifs [22–25], and to assign secondary structure [26].

A four-body statistical contact pseudopotential derived from Delaunay tessellation has been reported previously [13,14]. With this potential, the score of some particular amino acid quadruplet (i,j,k,l) is defined as

$$q_{ijkl} = \log_{10} \frac{f_{ijkl}}{c a_i a_j a_k a_l} \tag{1}$$

where f_{ijkl} is the observed frequency of simplices with amino acid types i,j,k, and l at their vertices in a large nonredundant training set S; a_i, a_j, a_k, and a_l are the observed frequencies of the individual amino acid types in S; and c is a combinatorial factor. Variations of this potential have been successfully applied to fold recognition [14,21] and the analysis of protein stability [15] and activity [20].

B. Properties of graphs and networks

A graph or network is a collection of nodes joined by edges [27,28]. In a weighted graph, an edge has a number weight associated with it denoting, for instance, physical distance between the nodes it connects or the relative difficulty in traversing the edge. In an unweighted graph, all edges are equivalent—one can arbitrarily assign them all weight 1. A directed graph has edges that can only be traversed in one direction. An undirected graph has edges that can be traversed in both directions. A connected graph is one in which it is possible to go between any pair of nodes via a path through a series of edges and other nodes. A completely connected graph is one where every node is directly connected by an edge to every other node. All networks considered here

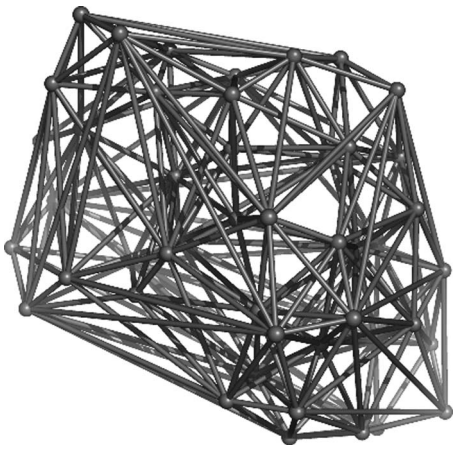

FIG. 1. Delaunay tessellation of crambin. Spheres $C\alpha$ atoms.

are undirected, unweighted, and connected. The order of a network is the number of nodes N it contains. The degree k of a node in an undirected graph is the number of edges impinging on it, and in a chemical context, this number is also called the coordination number. It should be noted that k will be used to refer to both the average degree of a whole network and also the degree of a single node. Context should make the meaning clear.

A minimum path between nodes i and j is one for which the sum of weights of the edges along the path is smallest from among all possible paths. The minimum path length L_{ij} between nodes i and j (also known as the chemical distance) is the sum of the weights along a minimum path. In the case described here, edges have weight 1, and a minimum path is one for which the fewest edges are traversed. The characteristic path length L of a network is the average of the minimum paths between all node pairs i,j, where $i \neq j$. The characteristic path length will also be referred to here as the mean minimum path length. In general, there are many paths between distinct nodes i and j that have the minimum path length. These paths are called geodesics [29].

Some classes of networks have the clustering property, which means that two nodes which are both joined by edges to a third, are more likely to also be joined to each other than are two nodes picked at random [30]. In such networks, there are well-defined neighborhoods—subsets of nodes tending to be connected to each other and tending not to be connected to other neighborhood subsets. The clustering coefficient of a node C_n is the number of actual edges E_n between neighbors of node n divided by the number of possible connections between those neighbors: $C_n = 2E_n/[k(k-1)]$, where k is the degree of node n. The clustering coefficient C for the entire network is the average of all the C_n.

Let $g_{ij}(v)$ be the number of geodesics between nodes i and j which pass through node v where $i \neq j \neq v$. Let g_{ij} be the number of all geodesics between i and j. Now define the ratio of the number of geodesics between i and j which pass through v to the total number of geodesics between i and j as $b_{ij}(v) = g_{ij}(v)/g_{ij}$. The betweenness centrality [29] of node v, denoted B_v, is then defined as

$$B_v = \frac{\sum_{i<j, j\neq v, i\neq v} b_{ij}(v)}{b_{\max}}, \quad \text{where} \tag{2}$$

$$b_{\max} = \frac{(N-1)(N-2)}{2} \quad \text{and } N = \text{number of nodes}.$$

The betweenness centrality is often abbreviated as the betweenness and measures the tendency of a node to be on minimum paths between other pairs of nodes. Note that $b_{\max}$ is a normalizing factor to keep betweenness in the range [0,1]. It is the maximum possible value of $b_{ij}(v)$, which occurs when the network is in a *star* configuration with v in the center surrounded by all the other nodes which are all connected to v, but not connected to each other [29]. The minimum value of $b_{ij}(v)$ is 0, as can happen for instance, when the degree of v is 1. Also note that betweenness is a property of an individual node, not a whole network.

Historically, graphs and networks have been divided into two extreme classes: regular and random [31,32]. In a regular network, each node has the same degree. Examples would be square or cubic lattices. A random graph is simply a collection of N nodes with edges connecting pairs of nodes with an independent probability p [33]. Such graphs have been very extensively studied. For example it is known that large random graphs have Poisson degree distributions [34]. For regular networks, $C = 3(k-2d)/4(k-d)$ and $L \sim N^{1/d}$ where k is the coordination number and d is the dimension in which the network exists [32,35]. Random networks have $C \sim k/N$ and $L \sim \ln(N)/\ln(k)$ [32]. In other words, regular graphs have comparatively large C and (at least for small d) large L, whereas sparse random graphs have small C and small L.

A class of networks, called small world networks, having properties of both regular and random networks has been characterized [30]. Small world networks have short path lengths with $L \sim \ln(N)/\ln(k)$ like random networks, but also have strong local clustering with $C \sim C_{\text{regular}}$. In other words, in a small world network there are well-defined neighborhoods where nodes tend to have mutual neighbors, like a regular graph, but one can traverse the graph in only a few moves as with a random graph. Such networks have since been found to be ubiquitous in nature [30,31]. Another class of networks of current interest, also ubiquitous in the real world, are *scale-free networks* [36] characterized by a degree distribution that follows a power law with exponent γ and, for some values of γ, a characteristic path length $L \sim \ln[\ln(N)]$ (ultrasmall world) [37].

C. Graph theoretic view of networks formed by inter-residue contacts in proteins

Here we model a protein structure as a network where nodes represent residues and unweighted edges represent the presence of putative interactions between residues. We call such networks residue contact networks or contact graphs and have analyzed a number of them, derived from real and model protein structures. Contact network analysis has been applied to molecular systems for a long time. Network analysis of molecules goes back at least to Wiener [38] who showed in 1947 a relationship between path length and the boiling points of paraffins. Recent papers have included detailed analysis of protein contact networks from a graph theoretic perspective. For instance, Vendruscolo *et al.* have

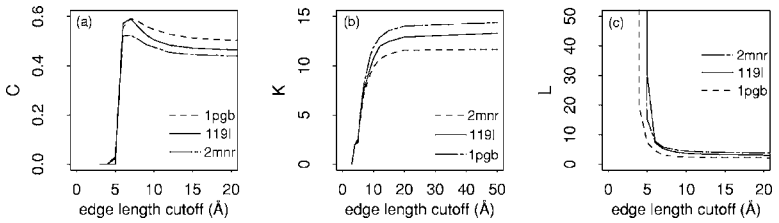

FIG. 2. The mean clustering coefficient, degree, and characteristic path length as a function of edge length cutoff for three representative PDB protein structures.

reported that networks formed by protein inter-residue contacts, where contact is defined as Cα separation $\leqslant$8.5 Å, have small world properties and that residues with high network betweenness make important contacts in model folding intermediates [8]. Using the characteristic path length of contact networks, Dokholyan *et al.* have reported being able to classify hypothetical "pretransition" and "post-transition" structures which could not be distinguished via rmsd deviation, solvent accessible area, or radius of gyration [9]. Atilgan *et al.* have shown that the average path length of a residue correlates with the amplitude of small fluctuations about the equilibrium position of that residue [39]. In this work we analyze residue contact networks where, instead of contact being defined by simple proximity, residues are considered to be in contact when a Delaunay edge joins them in the tessellation. Protein structures are stabilized in part by a large number of short range, non-covalent interactions and ideally graph edges should correspond to these real interactions. However, the tessellation of a protein chain can result in long simplex edges, particularly edges joining surface residues. It is sensible to exclude from our contact networks these "long edge" connections since they will not correspond to real physical interactions. Figures 2(a)–2(c) shows mean degree, characteristic path length, and clustering coefficient as a function of edge length cutoff for three representative sample structures. It is reasonable to choose a cutoff where the plots of all three have started to level off, which means a cutoff in the 8–12 Å range. We choose cutoffs of 8.5 and 10 Å for most of this work, but in some cases complement the data with a cutoff by the data without a cutoff since it results in simpler, more easily analyzed behavior.

It is important to recognize the distinction between the residue contact network and the Delaunay graph of the tessellated protein. The latter has edge lengths equal to the physical distances between residues. It is a geometric or spatial [28,32] graph, one which is by its nature explicitly understood to be embedded in a low-dimensional Euclidean space. In the case of a protein structure, the Euclidean space is obviously three dimensional and an example Delaunay graph is shown in Fig. 1. On the other hand, the contact graph is not a geometric graph but a relational graph [32] and is obtained from the Delaunay graph of the tessellated protein by setting all edge lengths equal to 1. Such a graph can also be embedded in a Euclidean space; however, a space of more than three (possibly significantly more than three) dimensions will be required in general [40].

II. EXPERIMENTAL METHODS

Several data sets of real proteins and artificial structures were compiled and Delaunay tessellated using the software developed by Zhibin Lu and the QHULL program [41]. The tessellated structures were then turned into networks or adjacency lists from which L, C, and B_v were calculated for further analysis. The tessellation software requires coordinates for all carbon α atoms (Cα's) without gaps. Therefore, in the case of artificial structures, only Cα coordinates were generated. In the case of real proteins, typically only about two thirds of Protein Data Bank (PDB) x-ray structures [42] can be Delaunay tessellated, mostly due to missing Cα coordinates.

The first data set consisted of the x-ray structures of 1364 nonhomologous protein chains obtained from the PISCES web server [43]. They had no gaps in the Cα coordinates, resolution $\leqslant$2.2Å, crystallographic R factor $\leqslant$0.23, and maximum pairwise sequence identity $\leqslant$30%. The chains from multimeric proteins were Delaunay tessellated in isolation from the other chains. This set will be abbreviated as 1364culled. The second was a set of 1364 simple computationally generated random polymers. Each consisted of a chain of N points with successive points separated by 3.83 units (typical Cα-Cα distance in real proteins in angstroms). The chain meandered in random directions subject to two constraints: it could not self-intersect and was confined to a sphere of diameter $7.177N^{1/3}+2$ in order to approximately match the size and shape of globular proteins. There was one such random polymer in the set for each real structure in 1364culled and it had the same number of residues as the real protein. This second set will be abbreviated as random strand. Obviously, unlike its corresponding real protein, each member of random strand had no secondary structure. A third data set consisted of 101 computationally generated, perfectly straight α helices with 5.4 units per turn, a helix diameter of 4.6 units, and a separation of monomer i and $i+1$ of 3.83 units. The lengths of the helices ranged from 50 to 550 residues. The third set will be abbreviated as artificial helix. A fourth set consisting of sparse, connected, random graphs was also generated. Again, there was one member of this set for each real structure in 1364culled and it had the same number of residues N as the real protein. Residue pairs were connected with probability p equal to the total number of edges in the real protein divided by $N(N-1)$. This set will be abbreviated as random network.

The first experiment consisted of computing and fitting C, L, and B_v for the data sets listed above. Second, Fourier spectra were computed on the betweenness profiles for structures from these data sets to look for periodicity, and identify possible closed loops. Last, C and L were computed for several native or decoy sets to determine if native and nonnative structures can be discriminated based solely on their contact network properties.

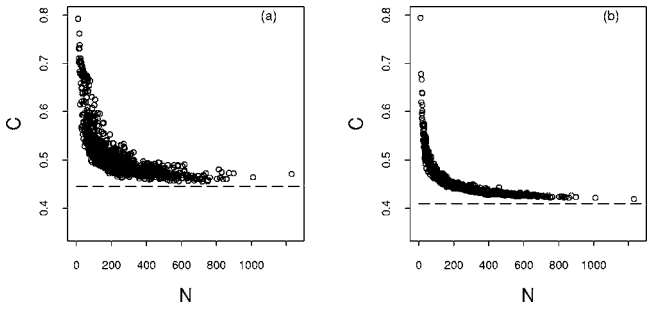

FIG. 3. Clustering coefficients for 1364 nonhomologous structures from PDB with and without 10 Å simplex edge length cutoff.

III. RESULTS

A. Are inter-residue contact networks small world?

We measured the mean clustering coefficient as a function of the number of residues for 1364culled [Figs. 3(a) and 3(b)]. The mean number of residues in 1364culled is about 230 and the mean degree (with 10 Å cutoff) is about 10. The observed mean clustering coefficient of 0.51 is significantly larger than that expected for a random graph, $k/N=0.043$, and is in the small world range. The plots for random strand are very similar to those of 1364culled. With the exception of the artificial-helix set with a 10 Å cutoff, for all tessellation derived contact networks considered here,C scaled roughly as $N^{-3/4}$. For the artificial-helix set with a 10 Å cutoff, C scaled as $1/N$. In all cases, C asymptotically approaches a nonzero value for large N [Figs. 3(a) and 3(b)]. The inverse scaling of C with N for roughly spherical point sets makes intuitive sense. A plot of N versus the ratio of the number of surface simplices (ones with unshared faces) to buried simplices looks similar to Figs. 3(a) and 3(b), flattening out at about $N=300$. A point on a surface face will not be surrounded by neighbors while a point in a buried simplex will. Neighbors on opposite sides of a point are unlikely to be neighbors of each other; therefore surface points will have higher clustering coefficients than buried points and the average C will be greater for point sets with a larger fraction of surface simplices. With a 10 Å cutoff, the inverse scaling of C with N for artificial helices is also understandable. In the interior of the helix each residue is in contact with the four residues that precede it and the four that follow and the connectivity patterns between these neighbors are always the same. The clustering coefficients C_{int} of all interior residues are therefore the same. The four residues at the N-terminal and C-terminal ends have fewer neighbors and they are more interconnected; hence their clustering coefficients C_{end} are larger. In the limit of infinite length, the average clustering coefficient for the whole helix will go to asymptotically to C_{int}.

The value of L as a function of the number of residues both with and without a 10 Å edge cutoff is shown in Figs. 4(a)–4(f) for 1364culled, random strand, and artificial helix. Note that there is a steeper line of data points above and to the left of the main body in the plot of L vs N in Fig. 4(a)—small proteins for which L scales differently than the rest. Such outliers are absent in the plot of L vs N with no cutoff and also absent in plots for random strand, which have no

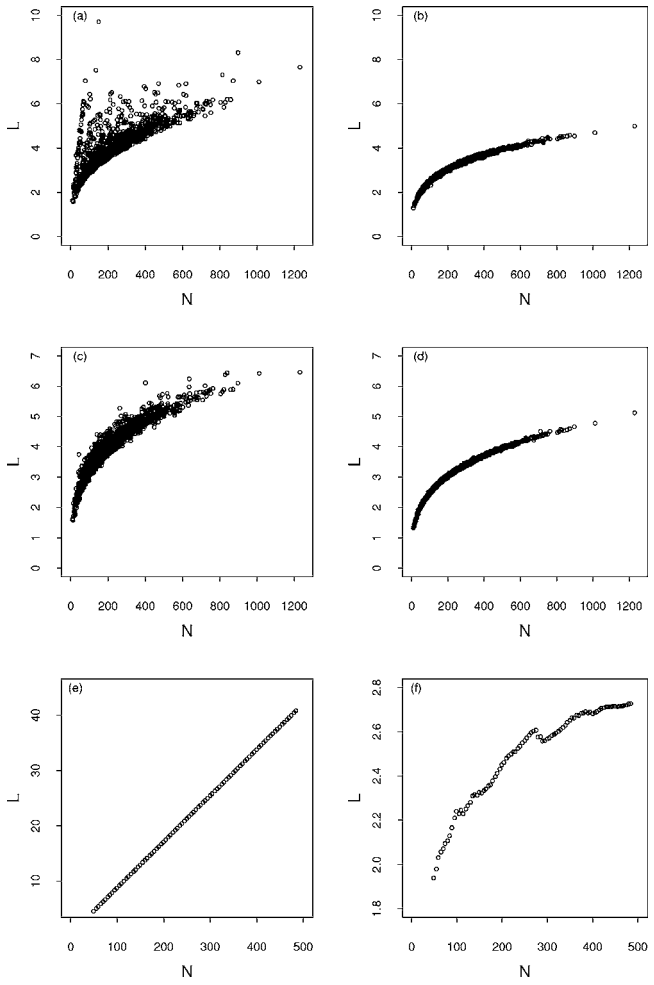

FIG. 4. Characteristic path length for contact graphs for 1364culled (top), random strand (middle), and artificial helix (bottom) with a 10 Å simplex edge length cutoff (left) and with no cutoff (right).

secondary structure. The steep line consists of small proteins, like, e.g., 1n7sC belonging to the SCOP [44] classes coiled coil and all α, that contain extended helices. The set of outliers between the main high density area of data and the steep line in the left part of 1364culled plot are mostly all α and all β, proteins. Structures from this set include, e.g., 1kx5B (all α with compact core and large extended loops), 1ospO (all β, not compact consisting of a big flat β sheet), and 1qfhA (all β, with two compact cores connected by a length of main chain).

The histogram of simplex edge lengths and inter-residue separation in primary sequence for random helix in Fig. 5 offers some explanation for the different trends in Figs. 4(a)–4(f). With a 10 Å cutoff, the ith residue in a long, straight helix is joined by edges to residues $i\pm1$, $i\pm2$, $i\pm3$, and $i\pm4$ and therefore helices are just one-lattices with $k=8$ (one-dimensional lattices with nodes connected to their four nearest neighbors on either side) [30]. Without the cutoff, the ith residue has additional connections to residues much further away in primary sequence. These long range edges act like the random rewiring in the Watts-Strogatz model [30] and the plots of L vs N are qualitatively different

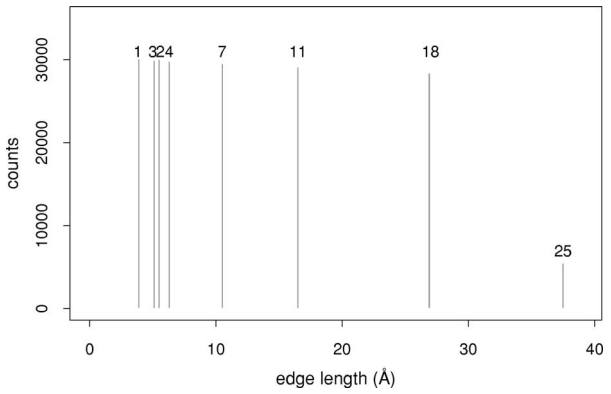

FIG. 5. Histogram of Delaunay simplex edge lengths in tessellations of random helix without edge length cutoff. Numbers above the bars are separation in primary sequence of the residues joined by a simplex edge of the length given on the x axis. Since the helices are perfectly regular repeating structures, the histogram consists of a few extremely narrow peaks, not a continuous distribution.

for the two cases: for $L_{\text{cutoff}} \sim N$ the networks behave like one-dimensional (1D) lattices whereas for $L_{\text{no cutoff}} \ll N$ the networks have small world character [Figs. 4(e) and 4(f)]. The network of contacts β sheets under a sufficiently short edge cutoff may approximate a 2D rectangular lattice and it is a reasonable conjecture that $L_{\text{cutoff}} \sim \sqrt{N}$; however, we have not systematically tested that hypothesis in this work. Large chains with compact, ellipsoidal shape and small compact chains with low secondary structure content do not display a sudden qualitative change in the characteristic path length with decreasing cutoff. A plot of L as a function of the number of residues, with residue-residue contact defined as simply Cα-Cα separation no greater than 8.5 Å instead of our Delaunay contact condition, looks very similar to Fig. 4(a). Therefore, regardless of the precise definition of contact, with commonly used contact cutoff distances there are at least two distinct trends in L as a function of N, which complicates analysis. In order to eliminate this complication we will analyze the results without cutoff then make a bounding argument to draw some conclusions on the scaling of L with a cutoff.

The observed path lengths for 1364culled fitted to $\ln(N)/\ln(k)$ (small world), $\sqrt[3]{N}$ (regular network in 3D), and $\sqrt[4]{N}$ for the data without edge cutoff and the resulting residuals are shown in Figs. 6(a)–6(f). All three models seem to produce good fits. However, for a legitimate linear least squares fit, residuals must be independent, homoscedastic, and normally distributed [45]. Quantile-quantile scatterplots of the residuals of the three fits versus a normal distribution (data not shown) indicate that all three sets of residuals are close to normal in distribution. But as the scatterplots in Figs. 6(b), 6(d), and 6(f) show, $\ln(N)/\ln(k)$ and $\sqrt[3]{N}$ have residuals with a systematic trend and nonuniform variance and therefore do not satisfy the regression requirements. No such systematic trend is seen in the residuals for $\sqrt[4]{N}$ and the vertical spread is fairly constant. The fit of observed L values to $\sqrt[4]{N}$ is therefore the best of the three and we can conclude that residue contact networks, though close in this regime, are not strictly small world as defined in the original paper

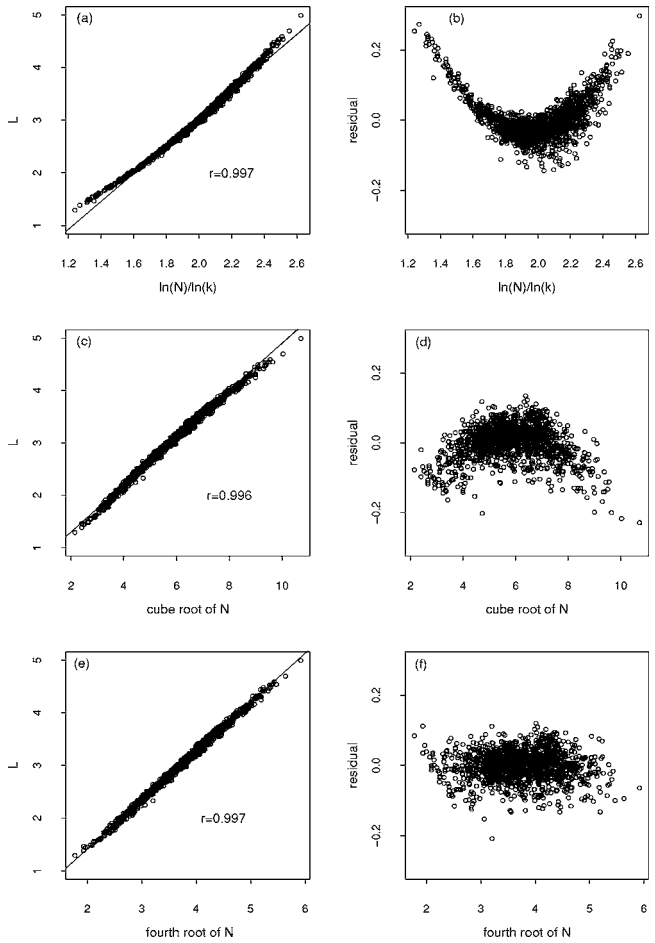

FIG. 6. Observed characteristic path lengths from 1364culled with no cutoff fitted to $\ln(N)/\ln(k)$ (a), $\sqrt[3]{N}$ (c), and $\sqrt[4]{N}$ (e) and the corresponding residuals [(b), (d), and (f)].

by Watts and Strogatz [30]. We already know, of course, that these networks are not regular, since not every residue has the same number of neighbors, and they are most definitely not scale-free networks [36], since the degree distribution has a well-defined peak and tails and since degrees vary by only about one order of magnitude (data not shown). Interestingly however, as mentioned before, C scales as $N^{-3/4}$, which is the same scaling as the preferential attachment model of scale-free networks [46]. L also scales as for random strand. However, with no cutoff L, scales as $\ln(N)/\ln(k)$ for artificial helix and the corresponding contact networks are true small world networks while with a 10 Å cutoff they are one-lattices.

In order to characterize the asymptotic behavior of L of the contact graphs of real proteins, we generated a second set of random self-avoiding model polymers which includes very large structures with up to 10 000 residues. Like the random-strands set they are computationally generated random self-avoiding polymers confined to a spherical volume of diameter $7.177(\sqrt[3]{N})+2$; however, the number of residues and the sequences do not correspond to real proteins. There is a strong correlation ($r=0.997$) between the path lengths L from 1364culled and L for the corresponding structures from random strand. Assuming that the correlation continues to

hold for $N>1200$, we can compute an approximate value of L for very large proteins, much larger than any real ones available in the PDB, by simply generating a suitably large random self-avoiding model polymer and computing its characteristic path length. We have done this for N in the range 100 to 10 000. The data were fitted as before and again analysis of residuals shows, that $L\sim\sqrt[4]{N}$ gives the best fit.

Returning to the case where we impose an edge cutoff, consider the ratio $R=L_{\text{prox cutoff}}/L_{\text{tess no cutoff}}$, where $L_{\text{prox cutoff}}$ is the characteristic path length of the contact network defined by simple proximity (Cα separation $\leqslant$8.5Å) and $L_{\text{tess no cutoff}}$ is the path length of the corresponding Delaunay tessellation derived contact graph with no edge cutoff. Analysis of the 1364culled data shows that the value of R varies considerably for small proteins, but in the limit of large N it tends toward about 1.5 and is never less than 1 for any value of N. We have therefore $L_{\text{prox cutoff}}\geqslant L_{\text{tess no cutoff}}$ and $L_{\text{tess no cutoff}}\sim O(\sqrt[4]{N})>O(\ln(N)/\ln(k))$. So it must follow that $L_{\text{prox cutoff}}>O(\ln(N)/\ln(k))$, therefore residue contact networks with cutoffs like those described by Vendruscolo *et al.* [8], are also not strictly small world. Finally, since the set of simplex edges connecting nodes in a residue contact graph under a cutoff is a subset of the set of edges with no cutoff, there cannot be a minimum path in the former that is not in the latter. It must therefore also be true that $L_{\text{tess cutoff}}\geqslant L_{\text{tess no cutoff}}\sim O(\sqrt[4]{N})>O(\ln(N)/\ln(k))$, where $L_{\text{tess cutoff}}$ is the path length of the Delaunay tessellation derived contact graph with an edge cutoff. Hence tessellation derived residue contact graphs under any cutoff cannot be strictly small world either.

This conclusion agrees with two previous observations. First, Newman has speculated that some networks classified as small world could be "almost regular" lattices in high dimensions [35]. With $L\sim N^{1/d}$ and $d>3$, L would be a slowly increasing function of N and with $C\gg k/N$ it would be difficult to distinguish from the small world case. And indeed we have shown in forthcoming work that Delaunay contanct networks under no cutoff are effectively four-dimensional objects, which is why $L\sim\sqrt[4]{N}$. Second, Watts has shown that only those contact graphs corresponding to geometric graphs with heavy right tailed edge length distributions can display small world behavior [32]. In other words, geometric graphs that have a relatively small fraction of edges with length on the order of the diameter D of the graph (including those with a fixed edge cutoff $r<D$) cannot have corresponding small world contact graphs. Even without an edge cutoff, the Delaunay graphs of 1364culled do not have such a heavy tailed edge length distribution (Fig. 7) and therefore the corresponding contact graphs should not be expected to be small world.

The distinction between small world and almost small world we have drawn here is probably not of practical importance for small N, as, for example, when N is the number of residues in a single globular protein chain. The contact networks will have small L and large C, i.e., small world character, so that they will be almost indistinguishable from true small world networks, which by definition must have a particular form for the scaling of L. However, if instead of constructing contact networks for residues, we constructed them for an all-atom representation of a protein or a solvated protein, N might conceivably go up by a large enough factor to show discernible non-small-world features.

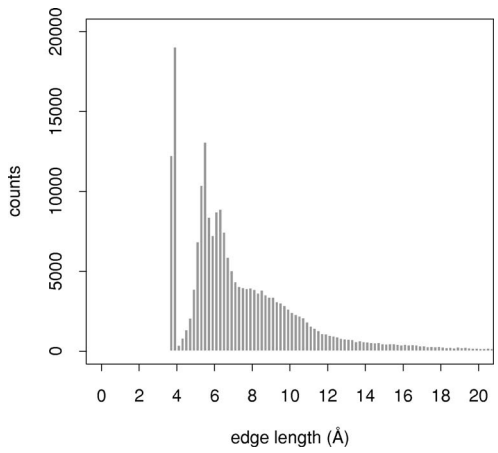

FIG. 7. Histogram of Delaunay edge lengths in tessellations of 1364culled without edge length cutoff.

B. Betweenness

Results reported earlier in the literature indicate that high betweenness residues may participate in important contacts. For example, Vendruscolo *et al.* [8] have reported that high betweenness residues play an important role in the folding nucleus of model transition state structures, and that some of these contacts are preserved in the native structure [8]. Shakhnovich *et al.* have identified "kinetically important" positions in a lattice model and a real structure by finding the positions with low sequence entropy in an alignment of a large number of sequences with high sequence-structure compatibility score when threaded onto the structures [47,48]—a method that does not use the network derived quantities we discuss here. Their model structure was a chain of 48 residues on a $3\times4\times4$ lattice and the real protein was chymotrypsin inhibitor 2 (CI2). The sequence-structure compatibility score was that of Miyazawa and Jernigan [49]. We have calculated betweenness profiles for both. In the real structure the contacts were defined using Delaunay tessellation, and since a rectangular lattice cannot be tessellated unambiguously, we simply define contacts of a site in the lattice model with its north, south, east, west, up, and down neighbors (no periodic boundary). The four residues in the model identified as kinetically important have the highest betweenness of all 48 in the model structure. Residues A35, I39, L68, I70, and I76, identified as kinetically important in CI2, have betweenness in the top 15% of all residues in the PDB structure 2ci2 under an 8.5 Å edge cutoff.

The betweenness of two selected structures calculated with a 10 Å cutoff is plotted in Figs. 8(a) and 8(c). Though noisy, the plots show some periodicity. Fourier power spectra for these structures were calculated and plotted in Figs. 8(b) and 8(d). Notice that the largest peaks correspond to a period of 20–40 residues. In order to study the properties of the betweenness for large numbers of structures, the following procedure was used.

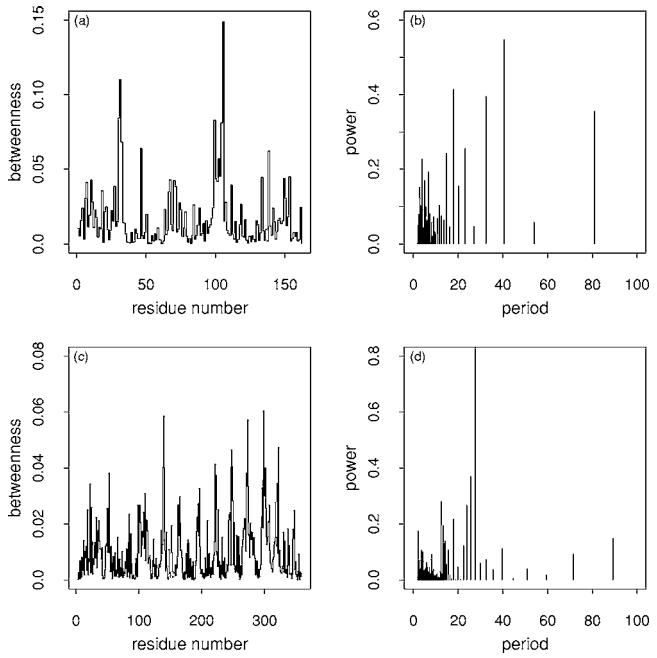

FIG. 8. Betweenness as a function of residue number and the Fourier spectrum of this betweenness profile in the Delaunay tessellations of the PDB protein structures 1191 (top) and 2mnr (bottom), both with a 10 Å edge cutoff.

(1) The set of structures was broken into six different categories according to length: 1–100, 101–200, 201–300, 301–400, 401–600, and 601–1200 residues.

(2) Power spectra for all structures in each group were generated.

(3) The power spectra were binned with a bin width of 0.1 residue.

(4) Since the peaks occur at noninteger values N/q, where $q=1,2,3\ldots$ and N is the sequence length and hence there are far more small period components than large period, all the binned spectra from a category were summed and then normalized by the number of terms in the sum for that bin; the resulting profile was then smoothed over a window of five residues.

Figures 9(a)–9(f) and 10(a)–10(f) show summed, binned, and smoothed spectra for the 1364culled and random-strand sets. Spectra for random network (not shown) do not have any well-defined peaks. Both the random-strand and 1364culled sets have well-pronounced peaks, albeit broad ones. However, the peak for random strand moves to the right with increasing length of the random polymer whereas for real structures it remains at about 30 residues.

The peak in the summed betweenness spectra is likely related to the peak at about 27 residues in the histogram of main chain separation for residue pairs with Cα-Cα distance less than a contact cutoff of 10 Å, reported by Berezovsky *et al.* [50–52] which is also invariant with respect to protein length. In this series of papers [50–52] the authors have described a peak at 22–32 residues in the histogram of sequence separation for Cα-Cα contact pairs. They define continuous stretches of 22–32 residues with the ends separated by less than 10 Å as closed loops. Their definition permits loops to overlap by up to five residues. If two putative loops

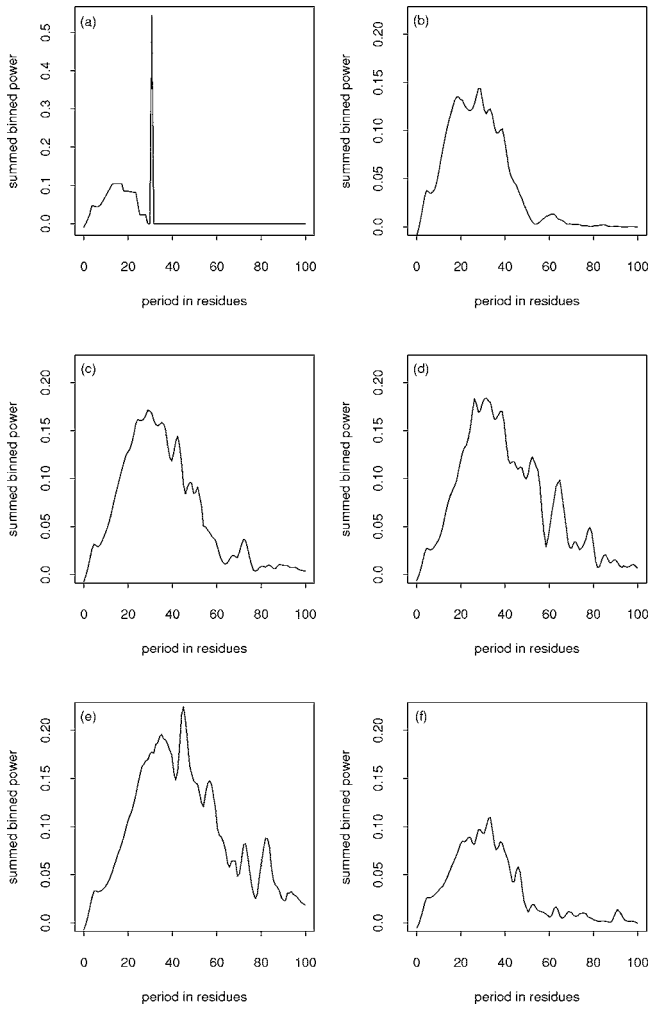

FIG. 9. The set of betweenness Fourier spectra from 1364culled was broken into six categories according to protein length: 1–100 residues (a), 101–200 residues (b), 201–300 residues (c), 301–400 residues (d), 401–600 residues (e), 601–1200 residues (f). The spectra were binned in units of 1/10 residue and then the power was summed for all proteins in that category for each of the six categories. As Fig. 8 shows, the Fourier peaks are not evenly spaced, so bins covering high frequency components will contain many more contributions than bins covering low frequencies. To compensate, the sum from each bin was then normalized by the number of different proteins with spectral components that contributed to the sum in that bin. Finally, the resulting six profiles were smoothed over a five residue window. The medians occur at periods of 17.7, 26.5, 33.0, 37, 42.9, and 30.6 residues for (a)–(f), respectively.

overlap by more than five residues, the one with the smaller Cα-Cα distance between its end residues is classified as a loop, the other is not. The loops typically cover about two-thirds of a protein sequence and are reported to have no preference for any class of secondary structure. Based on results from polymer physics, it has been suggested that 22–32 residues is the loop size that forms most readily in the unfolded state and such loops form first, creating folding nuclei which persist into the folded state [50–52].

Figures 11(a)–11(c) show the loops along with Kyte-Doolittle hydrophobicity [53], betweenness, and a four-body

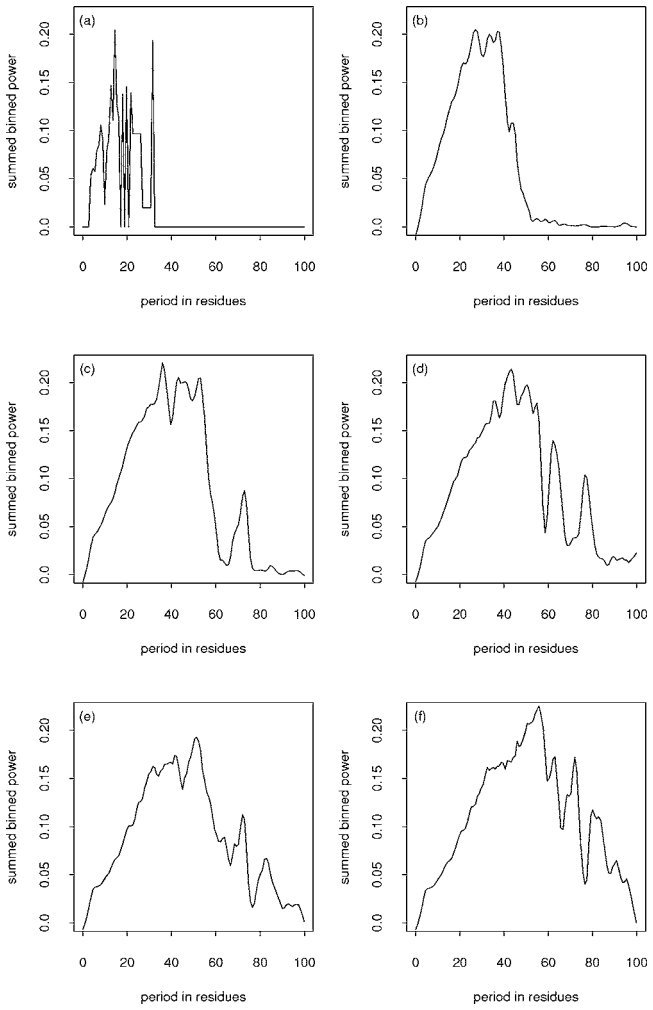

FIG. 10. The set of betweenness spectra from 1364 computationally generated random strands broken into six categories according to length. The spectra were binned, summed, and smoothed in the same way as in Fig 9. The medians occur at periods of 15.7, 28.5, 38.0, 42.7, 44.0, and 50.5 residues for (a)–(f), respectively.

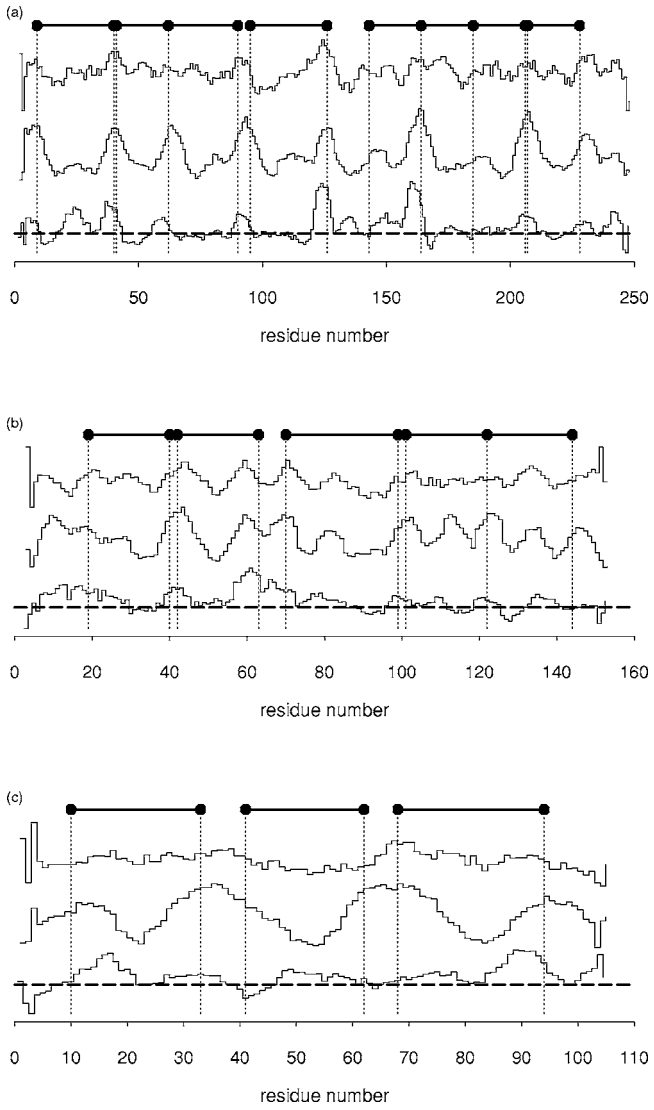

FIG. 11. Closed loops of Berezovsky *et al.* (dumbbells on top with loop end residues as circles) plotted with Kyte-Doolittle hydrophobicity (top profile), betweenness (middle profile), and tessellation based four-body 3D-1D sequence-structure compatibility profile (bottom profile) all smoothed over a seven-residue window. The plots are for the PDB protein chains 7timA (top), 1i1b (middle), and 256bA (bottom).

tessellation derived structure-sequence compatibility score [13] (described briefly in the introduction) all smoothed over a seven-residue window for three structures analyzed by Berezovsky *et al.* [50]. It is apparent from the plot that loop ends tend to occur at local maxima in betweenness and hydrophobicity and at positions where the four-body potential is positive (with this potential highly compatible sequence structure pairs have positive scores).

We have modified slightly the definition of loop introduced by Berezovsky *et al.* by (1) not allowing loop overlap for the sake of simplicity of analysis, and (2) requiring that a Delaunay simplex edge no longer than 10 Å joins two terminal residues which define the loop. The modified loops, which we will refer to as BGTD loops (for Berezovsky, Grosberg, Trifonov, and Delaunay), correspond well to the loops found under the unmodified definition in a number of representative structures analyzed previously [50]. We have compiled statistics on the 6536 BGTD loops in the 1364culled set to test the observations based on the small set of three structures shown in Figs. 11(a)–11(c). As previously reported, residues at the ends of the loops tend to be hydrophobic [52], but charged and polar residues also occur at loop ends and glycine, cysteine, tryptophan, and tyrosine are over represented if hydrophobicity were the sole factor (Table I). Table II shows that residues at the ends of these loops have betweenness and four-body potential values in the upper one-third of all residues in the protein. The mean Kyte-Doolittle hydrophobicity value of loop ends is also typically in the top 30% of all residues.

Figures 11(a)–11(c) show that in most cases the loop ends occur near but not necessarily exactly at local maxima of the betweenness plot. Therefore, we have computed the primary

TABLE I. Ratios of residue frequencies at *N*-term and *C*-term BGTD-loop ends to overall residue frequencies in 1364culled and correlations of these ratios with Kyte-Doolittle hydrophobicity scores.

	A	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*K*	*L*	*M*	*N*	*P*	*Q*	*R*	*S*	*T*	*V*	*W*	*Y*
N end	1.10	1.58	0.67	0.53	1.21	1.35	0.81	1.36	0.64	1.14	1.06	0.71	1.04	0.61	0.83	0.91	0.96	1.36	1.20	1.11
C end	1.23	1.63	0.56	0.47	1.34	1.29	0.93	1.26	0.60	1.18	1.25	0.87	0.89	0.67	0.77	0.89	0.95	1.28	1.13	1.23
Hydroph	1.8	2.5	−3.5	−3.5	2.8	−0.4	−3.2	4.5	−3.9	3.8	1.9	−3.5	−1.6	−3.5	−4.5	−0.8	−0.7	4.2	−0.9	−1.3
Correlation of *N* term loop end frequencies with Kyte-Doolittle hydrophobicity scores: 0.823																				
Correlation of *C* term loop end frequencies with Kyte-Doolittle hydrophobicity scores: 0.817																				

sequence separation between each residue in 1364culled and the position with the maximum values of betweenness and potential in a window of size 21 centered on the residue. Table II shows that loop ends tend to be closer to such local maxima than about two-thirds of all residues. The high potential score of loop ends indicates they are important to protein stability [21] and the high betweenness at the ends of ubiquitous ~30 residue loops explains the persistent peaks at ~30 residues in the summed Fourier spectra. Figures 11(a)–11(c) also indicate that hydrophobicity, four-body potential, and betweenness tend to have local maxima in the same places. However, the pairwise correlations between them are not strong (Table III), so it can be suggested there is information in each profile not contained in the others.

To analyze the correlations in the betweenness values for the residues in contact, the residues from four structures analyzed by Berezovsky *et al.* [50] were divided into three equal groups of high, intermediate, and low betweenness. The edges joining residues from different combinations of groups were then counted (high-low, high-intermediate, high-high, etc.). Table IV shows that high and low betweenness residues preferentially form contacts within their respective groups.

For the structure 16pk, Fig. 12 shows the contacts defining all the overlapping putative loops, of which the BGTD loops are a subset. Notice these contacts are not uniformly distributed along the sequence, and this is typical of other structures as well. Instead, there are pairs of contiguous groups of residues separated by ~30 with a large number of contacts between the two groups. For example, the groups (61–66) and (89–92) have a total of ten contacts between them and the groups (99–114) and (128–137) have a total of eighteen. In our modified loop definition presented above as well as in [50], such closed loops have been characterized by a single contact, namely, the one with the smallest Cα-Cα distance. However, the difference in Cα-Cα distance between the smallest and second smallest contacts in the two loops considered above is less than half an angstrom. The choice between overlapping loops with contact distances that differ by such a small amount is somewhat arbitrary, and as Fig. 12 shows choosing one contact over another can shift the position of the closed loop in the sequence by ten residues or more. It might be better to define closed loops by a group of contacts or by some aggregate measure of connectivity such as network centrality. For example one could choose a closed loop from among overlapping candidates by picking the one defined by a contact between a pair of residues with the highest betweenness instead of the pair with the smallest physical separation.

To summarize, our analysis indicates that globular proteins have a number of high betweenness residues which, also have high values of a residual four-body knowledge-based potential, and are likely to play an important structural role. These residues tend to preferentially form contacts with each other and some of them tend to form the stems of ~30 residue closed loops, which are known to play an important role in forming a folding nucleus. The tessellation approach can be used to define and analyze these substructures in a robust quantitative manner.

C. Discrimination between native and decoys with graph methods

Using the characteristic path length of contact networks, Dokholyan *et al.* were able to distinguish pretransition and

TABLE II. Mean betweenness and potential quantiles of ends of nonoverlapping loops of length 22–32 residues closed by the shortest Delaunay edge (BGTD loop) and of all putative loops of length 22–32 residues closed by Delaunay edges up to 10 Å for 1364culled. Betweenness and four-body KB potential are unsmoothed. The distance to max is the distance from the loop end residue to the maximum value in a 21-residue window centered at the loop end. Loop ends therefore tend to have betweenness and potential in about the top one-third of all residues and are closer to local maxima in betweenness and potential than about two-thirds of all residues.

	Betweenness quantile	Distance to max quantile	Potential quantile	Distance to max quantile
BGTD-loop ends	69.5	35.1	66.6	40.1
Ends of all putattive closed Loops of 22–32 residues	69.3	36.9	67.5	40.6

TABLE III. Correlations between four-body knowledge-based potential, Kyte-Doolittle hydrophobicity, and network properties for all residues in 1364culled (betweenness, knowledge-based potential, and hydrophobicity smoothed over a seven-residue window).

	Clust. coef.	Hydrophobicity	Degree	Path length	Potential
Betweeness	−0.443	0.172	0.429	−0.418	0.196
Clust. coef.		−0.270	−0.868	0.259	−0.201
Hydrophobicity			0.335	−0.112	0.332
Degree				−0.271	0.256
Path length					−0.0987

post-transition structures—hypothetical folding intermediates that live along the reaction coordinate on either the nonnative or native side of the transition state [9]. An obvious related question is whether native or close to native conformations have different contact network properties than misfolded conformations.

To test this, we computed the characteristic path length L, mean clustering coefficient C, and mean degree K for a set of proteins for which structure decoys are available. We have used structures from four sets of the multiple decoy section of the Decoys'R Us website [54]. Decoys from the hgstructal sets are globins and have been built by comparative

TABLE IV. The counts of Delaunay edges (10 Å cutoff) connecting high, intermediate, and low betweenness residues for four selected structures tested against a null hypothesis with χ^2 tests. High betweenness residues tend to have a higher degree than medium, which in turn tend to have a higher degree than low betweenness residues. Therefore a null with all interclass contacts equally probable is not adequate. The fraction of edges impinging on each class is computed from the real tessellated structure. The expected frequencies in the null are then taken to be the products of these fractions. In parentheses are the factors by which the actual counts differ from the null.

	1i1b (151 residues, 801 simplex edges)		
	Top 33%	Middle 33%	Bottom 33%
Top 33%		196 (1.31)213 (0.93)	90 (0.56)
Middle 33%		81 (0.92)	157 (1.27)
Bottom 33%			64 (1.45)
	$\chi^2=67.55$, p value $\leqslant 7.48\times10^{-14}$		
	1thbA (141 residues, 750 simplex edges)		
	Top 33%	Middle 33%	Bottom 33%
Top 33%	177 (1.44)	189 (0.91)	65 (0.42)
Middle 33%		89 (1.02)	144 (1.11)
Bottom 33%			86 (1.79)
	$\chi^2=109.48$, p value $\leqslant 2.2\times10^{-16}$		
	256bA (105 residues, 550 simplex edges)		
	Top 33%	Middle 33%	Bottom 33%
Top 33%	127 (1.38)	147 (1.02)	49 (0.40)
Middle 33%		61 (1.09)	83 (0.86)
Bottom 33%			83 (2.08)
	$\chi^2=107.47$, p value $\leqslant 2.2\times10^{-16}$		
	7timA (247 residues, 1437 simplex edges)		
	Top 33%	Middle 33%	Bottom 33%
Top 33%	371 (1.47)	329 (0.81)	136 (0.46)
Middle 33%		185 (1.14)	268 (1.14)
Bottom 33%			148 (1.74)
	$\chi^2=209.83$, p value $\leqslant 2.2\times10^{-16}$		

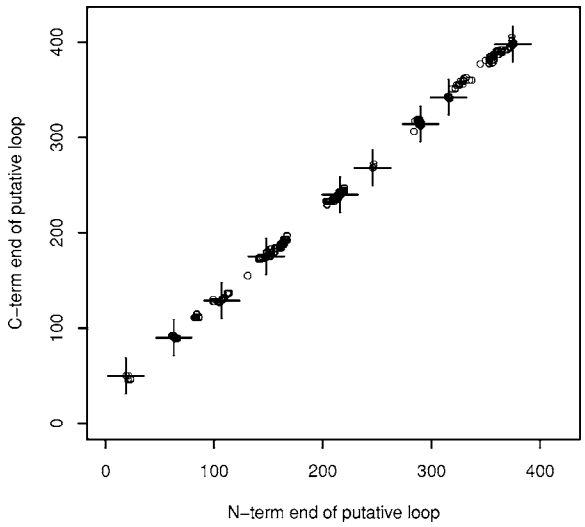

FIG. 12. All Delaunay contacts between residues separated by 22–32 in primary sequence and with edge lengths $\leqslant 10$ Å for the PDB structure 16pk. The ends of BGTD loops are denoted with crosses.

modeling using other globins as templates with the program segmod [54,55]. Decoys from the fisa set are small, α-helical structures constructed from fragments of unrelated protein structures with similar local sequences using a simulated annealing procedure [56]. Decoys from the lmds set were derived from the experimental secondary structures of ten small proteins that belong to diverse structural classes. Each decoy is at a local minimum of a "handmade" energy function [54]. The 4state-reduced set contains decoys for small proteins and the Cα positions for these decoys were generated by exhaustively enumerating ten selectively chosen residues in each protein using a four-state off-lattice model [57].

Table V shows the rank of the native in a sorted list of the native plus decoys for C, L, and K under three edge length cutoffs. One can see that certain of these network parameters can, to some extent, discriminate native from decoy. For example, for the 4state-reduced set, $C_{6.5}$, the mean clustering coefficient under a 6.5 Å cutoff, of the native tends to be greater than that of the decoys as does L_{inf} the mean characteristic path length with no cutoff. For the fisa set, $C_{8.5}$, $L_{6.5}$, and $L_{8.5}$ all strongly tend to be lower for native than for decoys. For the lmds set, $C_{8.5}$ and $L_{6.5}$ tend to be lower for the native and $K_{6.5}$ and $K_{8.5}$ tend to be higher. For hg-structal, $C_{8.5}$, $L_{6.5}$, and K_{inf} all tend to be lower for native than for decoys, but the trend is not as strong as with the other three sets.

Because these decoy sets were constructed by quite different means, it is not unreasonable that the network properties of decoys relative to the native should be different for different sets. It is interesting to note that C, L, and K seem to have the most discriminatory power with either no cutoff or a short cutoff (6.5–8.5 Å). Cutoffs in the intermediate range (10–15 Å) do not work as well (data not shown). With no cutoff, connections between widely separated surface residue are included in the contact network. With a very short cutoff, the network consists mostly of connections between residues in the core, which may shed some light on the trends in Table V. For example, native structures in the lmds sets tend to have high $K_{8.5}$ and low $C_{8.5}$ which means core residues in the natives tend to be surrounded by more neighbors than in decoys.

For the majority of structures, the native is systematically in the top or bottom 50% of the decoy/native set sorted on one of these discriminatory network parameters. Therefore one can usually eliminate half the decoys as "non-native" based on contact network criteria with no reference at all to an amino acid dependent energy function. One can be more restrictive by paring the set of decoys down to the short list of structures in the top or bottom 50% for both of two discriminatory parameters. For example, Table V shows that of the 631 structures in the 1ctf 4state-reduced decoy set, 191 (30%) of them have both $C_{6.5}$, and L_{inf} in the top 50th percentile and that the native is one of these structures. For the hg-structal set, such a two-part test does not work as well as with the other sets, but still succeeds in 12 out of 26 cases, significantly above ($p < 0.05$) the expected number of 6.5 successes given the null that the native is equally likely to have high or low rank in a sorted list, and that the pair of network parameters are independent (which is not strictly true from Table III).

The four Decoy's R Us sets used here are too small and too biased toward small and α-helical proteins to draw any strong general conclusions; however, our results suggest that native protein structures may have specific contact network properties that can be used to distinguish them from many decoys. It should also be noted that one can use the geometric properties of the tessellations of decoy and native structures, still ignoring the amino acid sequence, to discriminate decoy from native more effectively than with contact network parameters (this work is currently in preparation for publication).

IV. CONCLUSIONS

We have applied graph theory and the theory of complex networks to the analysis of various protein and proteinlike models where the polymer is represented by a network of contacts between the residues defined through the Delaunay tessellation. We have shown that these protein contact networks have small world character, but in several respects deviate from strictly defined small world networks. Like the small world networks, protein contact networks are highly clustered and the average graph distance between the nodes is short; however, unlike the small world networks, they have too few long edges that span the entire graph. In addition, our results suggest that networks formed by native structures and by most misfolded decoys have different graph properties, and this can be used as a simple and efficient sequence-independent filter for the discrimination between the native-like folds and decoys. Closed loops, consisting of 22–32 consecutive residues, where terminal residues are Delaunay neighbors are detected in tessellated protein structures. Locations and lengths of these loops correlate well with those

TABLE V. The rank of the native in a sorted list of the native plus decoys for C, L, and K under three edge length cutoffs for each decoy set. Quantities in bold are used together in a two-part test to generate a short list of more nativelike structures. For the 501 structures in the 1hdd:C set, for example, the intersection of the subset of structures with mean C under an 8.5 Å cutoff in the lower 50th percentile with the subset of structures with mean L under a 6.5 Å cutoff in the lower 50th percentile gives a short list of 134 structures and the native is in this short list.

PDB ID	SCOP class	Length	Decoy set	Number decoys	Range of rms deviation	C (6.5 Å) rank	C (8.5 Å) rank	C_{inf} rank	L (6.5 Å) rank	L (8.5 Å) rank	L^{inf} rank	K (6.5 Å) rank	K (8.5 Å) rank	K_{inf} rank	Size short list	Native in list?
1ctf	d	68	4-state	631	1.3–9.1	**1**	23	49	199	137	**70**	233	410	383	191	Yes
1r69	a	63	4-state	676	0.9–8.3	**37**	227	227	632	489	**215**	89	295	345	184	Yes
1sn3	g	65	4-state	661	1.3–9.1	**77**	379	644	560	524	**41**	61	179	551	170	Yes
3icb	a	75	4-state	654	0.9–9.4	**65**	138	62	168	464	**234**	466	399	208	189	Yes
4pti	g	58	4-state	688	1.4–9.3	**248**	369	466	583	513	**508**	303	417	307	190	No
4rxn	g	54	4-state	678	1.4–8.1	**20**	47	88	190	201	**54**	312	481	603	185	Yes
1fc2:C	a	42	fisa	501	3.1–10.6	484	**501**	172	**459**	491	128	405	7	111	149	Yes
1hdd:C	a	56	fisa	501	2.8–12.9	105	**415**	363	**465**	490	369	101	28	60	134	Yes
2cro	a	64	fisa	501	4.3–12.6	131	**477**	74	**475**	496	44	10	5	338	157	Yes
4icb	a	76	fisa	501	4.8–14.1	141	**501**	351	**501**	501	24	6	1	316	163	Yes
1ash	a	147	hg-structal	30	2.2–6.9	27	**18**	19	**29**	25	15	6	15	13	9	Yes
1bab:B	a	145	hg-structal	30	0.7–6.9	12	**22**	22	**21**	22	6	10	8	23	9	Yes
1col:A	f	197	hg-structal	30	12.4–30.3	24	**30**	12	**30**	30	1	1	1	30	12	Yes
1ecd	a	136	hg-structal	30	1.5–6.2	7	**19**	1	**28**	9	11	7	15	17	9	Yes
1emy	a	153	hg-structal	30	0.7–9.3	3	**17**	17	**26**	13	8	6	18	11	8	Yes
1flp	a	142	hg-structal	30	1.7–7.2	9	**16**	4	**15**	19	3	24	12	29	10	No
1gdm	a	153	hg-structal	30	2.6–8.4	26	**16**	26	**11**	3	30	1	10	1	8	No
1hbg	a	147	hg-structal	30	2.1–6.9	23	**21**	19	**17**	10	24	2	9	9	8	Yes
1hbh:A	a	141	hg-structal	30	1.0–6.3	8	**14**	1	**21**	22	18	8	5	13	8	No
1hbh:B	a	146	hg-structal	30	1.0–7.3	27	**6**	17	**10**	6	26	21	22	4	9	No
1hda:A	a	141	hg-structal	30	0.5–5.8	9	**19**	14	**26**	20	22	16	19	26	8	Yes
1hda:B	a	145	hg-structal	30	0.5–5.6	20	**12**	12	**8**	22	9	13	6	18	10	No
1hlb	a	157	hg-structal	30	2.9–7.0	2	**6**	1	**6**	5	17	30	29	26	9	No
1hlm	a	158	hg-structal	30	3.0–8.7	26	**15**	23	**4**	6	24	30	30	16	10	No
1hsy	a	153	hg-structal	30	0.8–9.7	5	**8**	10	**17**	10	19	9	24	9	7	No
1ith:A	a	141	hg-structal	30	1.6–6.1	24	**25**	1	**23**	20	27	20	3	7	8	Yes
1lht	a	153	hg-structal	30	0.8–9.7	8	**18**	29	**10**	13	8	11	17	18	6	No
1mba	a	146	hg-structal	30	1.8–7.3	8	**27**	22	**7**	7	9	8	9	28	8	No
1mbs	a	153	hg-structal	30	1.7–9.3	12	**7**	12	**26**	14	4	11	18	21	10	No
1myg:A	a	153	hg-structal	30	0.5—9.6	5	**22**	3	**24**	15	14	6	16	17	8	Yes
1myj:A	a	153	hg-structal	30	0.6–7.9	1	**21**	19	**19**	14	15	7	11	21	9	Yes
2dhb:A	a	141	hg-structal	30	0.6–6.4	16	**26**	18	**12**	21	18	23	19	28	9	No
2dhb:B	a	146	hg-structal	30	0.9–7.1	20	**10**	21	**12**	7	11	23	24	21	9	No
2pgh:A	a	141	hg-structal	30	0.7–6.5	22	**25**	13	**23**	13	26	18	21	25	9	Yes
2pgh:B	a	146	hg-structal	30	0.8–7.5	30	**9**	20	**21**	25	27	18	10	5	8	No
4sdh:A	a	145	hg-structal	30	2.3–6.4	25	**30**	10	**24**	30	12	4	1	24	10	Yes
1dtk	g	57	lmds	216	4.3–12.6	87	**169**	5	**211**	208	1	2	1	214	59	Yes
1fc2:C	a	43	lmds	501	4.0–8.4	500	**496**	25	**498**	498	36	134	1	345	155	Yes
1igd	d	61	lmds	501	3.1–12.6	249	**460**	302	**337**	179	171	89	214	312	151	Yes
1shf:A	b	59	lmds	438	4.4–12.3	13	**101**	276	**435**	285	25	2	174	426	113	No
2cro	a	65	lmds	501	3.9–13.5	383	**501**	47	**501**	501	38	3	1	307	154	Yes
2ovo	g	56	lmds	348	4.4–13.4	230	**311**	147	**269**	249	119	39	170	251	96	Yes
4pti	g	58	lmds	344	4.9–13.2	217	**320**	250	**343**	333	208	1	7	198	97	Yes

defined by other means, however the proposed technique affords a more consistent approach to loop identification. The systematic analysis of the locations, geometric and compositional features of these loops may provide important insights into the protein folding and structure.

ACKNOWLEDGMENTS

We wish to thank Zhibin Lu for writing part of the codes used in calculations and Greg Reck for helpful discussion on this work.

[1] M. Vendruscolo and E. Domany, Vitam. Horm. (San Diego, CA, U. S.) **58**, 171 (2000).
[2] U. Bastolla *et al.*, Proteins **58**, 22 (2005).
[3] M. Porto *et al.*, Phys. Rev. Lett. **92**, 218101 (2004).
[4] G. Amitai *et al.*, J. Mol. Biol. **344**, 1135 (2004).
[5] A. del Sol, H. Fujihashi, and P. O'Meara, Bioinformatics **21**, 1311 (2005).
[6] A. del Sol and P. O'Meara, Proteins **58**, 672 (2005).
[7] L. H. Greene and V. A. Higman, J. Mol. Biol. **334**, 781 (2003).
[8] M. Vendruscolo *et al.*, Phys. Rev. E **65**, 061910 (2002).
[9] N. V. Dokholyan *et al.*, Proc. Natl. Acad. Sci. U.S.A. **99**, 8637 (2002).
[10] A. Poupon, Curr. Opin. Struct. Biol. **14**, 233 (2004).
[11] I. Vaisman, in *Handbook of Computational Statistics* (Springer, New York, 2004), p. 981.
[12] A. Okabe, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams* (Wiley, Chichester, 2000).
[13] R. K. Singh, A. Tropsha, and I. I. Vaisman, J. Comput. Biol. **3**, 213 (1996).
[14] A. Tropsha *et al.*, Pac. Symp. Biocomput. 614 (1996).
[15] B. Krishnamoorthy and A. Tropsha, Bioinformatics **19**, 1540 (2003).
[16] V. A. Ilyin, A. Abyzov, and C. M. Leslin, Protein Sci. **13**, 1865 (2004).
[17] J. Roach *et al.*, Proteins **60**, 66 (2005).
[18] D. L. Bostick, M. Shen, and I. I. Vaisman, Proteins **56**, 487 (2004).
[19] J. Liang *et al.*, Proteins **33**, 1 (1998).
[20] M. Masso and I. I. Vaisman, Biochem. Biophys. Res. Commun. **305**, 322 (2003).
[21] C. W. Carter, Jr. *et al.*, J. Mol. Biol. **311**, 625 (2001).
[22] A. Tropsha *et al.*, Methods Enzymol. **374**, 509 (2003).
[23] S. A. Cammer, R. P. Carty, and A. Tropsha, in *Proceedings of the 3rd International Workshop on Algorithms for Macromolecular Modeling*, edited by T. Schlick and H. H. Gan (Springer, New York, 2000), p. 477.
[24] H. Wako and T. Yamato, Protein Eng. **11**, 981 (1998).
[25] J. Huan *et al.*, Pac. Symp. Biocomput. 411 (2004).
[26] T. Taylor *et al.*, Proteins **60**, 513 (2005).
[27] B. Bollobás, *Graph Theory: An Introductory Course* (Springer-Verlag, New York, 1979).
[28] D. B. West, *Introduction to Graph Theory* (Prentice-Hall, Upper Saddle River, NJ, 2001).
[29] L. Freeman, Sociometry **40**, 35 (1977).
[30] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[31] R. Albert, H. Jeong, and A. Barabasi, Nature (London) **401**, 130 (1999).
[32] D. J. Watts, *Small Worlds: The Dynamics of Networks Between Order and Randomness* (Princeton University Press, Princeton, NJ, 1999).
[33] P. Erdos and A. Renyi, Publ. Math. (Debrecen) **6**, 290 (1959).
[34] M. Newman, D. Watts, and S. Strogatz, Proc. Natl. Acad. Sci. U.S.A. **99**, 2566 (2002).
[35] M. Newman, J. Stat. Phys. **101**, 819 (2000).
[36] L. A. Amaral *et al.*, Proc. Natl. Acad. Sci. U.S.A. **97**, 11149 (2000).
[37] R. Cohen and S. Havlin, Phys. Rev. Lett. **90**, 058701 (2003).
[38] H. Wiener, J. Am. Chem. Soc. **69**, 17 (1947).
[39] A. R. Atilgan, P. Akan, and C. Baysal, Biophys. J. **86**, 85 (2004).
[40] P. Erdos, F. Harary, and W. T. Tutte, Mathematika **12**, 118 (1965).
[41] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, ACM Trans. Math. Softw. **22**, 469 (1996).
[42] H. M. Berman *et al.*, Acta Crystallogr., Sect. D: Biol. Crystallogr. **58**, 899 (2002).
[43] G. Wang and R. L. Dunbrack, Jr., Bioinformatics **19**, 1589 (2003).
[44] A. G. Murzin *et al.*, J. Mol. Biol. **247**, 536 (1995).
[45] G. E. P. Box and D. R. Cox, J. R. Stat. Soc. Ser. B. Methodol. **26**, 211 (1964).
[46] E. Ravasz and A. L. Barabasi, Phys. Rev. E **67**, 026112 (2003).
[47] E. Shakhnovich, V. Abkevich, and O. Ptitsyn, Nature (London) **379**, 96 (1996).
[48] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[49] S. Miyazawa and R. Jernigan, Macromolecules **18**, 534 (1985).
[50] I. N. Berezovsky, A. Y. Grosberg, and E. N. Trifonov, FEBS Lett. **466**, 283 (2000).
[51] I. N. Berezovsky *et al.*, Protein Eng. **15**, 955 (2002).
[52] I. N. Berezovsky *et al.*, Proteins **45**, 346 (2001).
[53] J. Kyte and R. F. Doolittle, J. Mol. Biol. **157**, 105 (1982).
[54] R. Samudrala and M. Levitt, Protein Sci. **9**, 1399 (2000); http://dd.stanford.edu
[55] M. Levitt, J. Mol. Biol. **226**, 507 (1992).
[56] K. T. Simons *et al.*, J. Mol. Biol. **268**, 209 (1997).
[57] B. Park and M. Levitt, J. Mol. Biol. **258**, 367 (1996).