Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 365410, 9 pages
http://dx.doi.org/10.1155/2013/365410

Research Article

The Effect of Edge Definition of Complex Networks on Protein Structure Identification

Jing Sun,[1] Runyu Jing,[1] Di Wu,[1] Tuanfei Zhu,[2] Menglong Li,[1] and Yizhou Li[1]

[1] *College of Chemistry, Sichuan University, Chengdu 610064, China*
[2] *College of Computer Science, Sichuan University, Chengdu 610064, China*

Correspondence should be addressed to Menglong Li; liml@scu.edu.cn and Yizhou Li; liyizhou_415@163.com

Received 6 December 2012; Revised 23 January 2013; Accepted 25 January 2013

Academic Editor: Guang Hu

The main objective of this study is to explore the contribution of complex network together with its different definitions of vertexes and edges to describe the structure of proteins. Protein folds into a specific conformation for its function depending on interactions between residues. Consequently, in many studies, a protein structure was treated as a complex system comprised of individual components residues, and edges were interactions between residues. What is the proper time for representing a protein structure as a network? To confirm the effect of different definitions of vertexes and edges in constructing the amino acid interaction networks, protein domains and the structural unit of proteins were described using this method. The identification performance of 2847 proteins with domain/domains proved that the structure of proteins was described well when R_{C_α} was around 5.0–7.5 Å, and the optimal cutoff value for constructing the protein structure networks was 5.0 Å (C_α-C_α distances) while the ideal community division method was community structure detection based on edge betweenness in this study.

1. Introduction

Protein structure comparison and classification are a difficult but important task since structure is a determinant for molecular interaction and function [1]. Protein folds into a specific conformation for its function depending on interactions between residues. Consequently, a protein structure can be treated as a complex system comprised of individual components residues. The method of complex network has been widely applied in various types of fields such as disease [2–4], drug target [5], drug design [6]. Network analysis facilitates the characterization of such complex system and its individual components [7, 8]. This provides novel insights into understanding the protein folding mechanism [9, 10], stability [11], function [9, 12, 13], and dynamics [14] and, more specifically, the study of protein structures. Viewing the protein structure as the an intricate network of interacting residues, metastructure analysis was proved to be an effective tool for large-scale (genome-wide) protein sequence analysis target selection for structural genomics and the identification of intrinsically unstructured (unfolded) proteins [15]. Analysis of the protein structure graphs showed that the aromatic residues along with arginine, histidine, and methionine act as strong hubs at high interaction cutoffs, which are found to play a role in bringing together different secondary structural elements in the tertiary structure of the proteins [11]. Through transforming the protein structure into residue interaction graphs, active site, ligand-binding, and evolutionary conserved residues were found to have high closeness values typically. This property will then be used to identify key protein residues [16]. Moreover, software tools were presented for the automatized generation, 2D visualization, and interactive analysis of residue interaction networks, which proved that residue networks are crucial for understanding structure-function relationships [17]. A novel web server, RING, was presented to construct physicochemically valid residue interaction networks interactively from PDB files for subsequent visualization in the Cytoscape platform [18]. The application of Cytoscape plug-ins, NetworkAnalyzer [19], and RINalyzer [17] were demonstrated for the standard and advanced analyses of network topologies [20].

In these studies, different strategies were used to define a vertex in literature: (a) only the C_α [9, 10, 15, 21–23] or C_β [21, 24] of an amino acid; (b) the center of the side

chain [11]; (c) all atoms in a residue were taken into account [16, 25]. Moreover, definition of edge also appears crucial in the construction of such networks. The characterization of protein structure is sensitive to the threshold for edges such as 5 Å (distances between two atoms from two amino acid residues) [25], 8 Å (C_α-C_α distances) [15], 8.5 Å (pairs of amino acids) [9], and a strict cutoff value of 7 Å [9, 10, 15, 21–23] based on the discovery that representing amino acids by C_α atoms may introduce bias for cutoffs below 6.8 Å [23].

Which strategy is more reasonable among all these choices? Studies have been made to find the answer. Three models were compared to prove the effects of the anisotropic nature of the side chain on the identification of the contact amino acid pairs [26]. The main objective of this study is to explore the contribution of complex network together with its different definitions of vertexes and edges to describing the structure of proteins. Automatic decomposition of protein structures into domains remains a challenging problem [27], and numbers of computer algorithms have been proposed [27–30]. Since domains can be considered as semi-independent structural units of a protein capable of folding independently [31, 32], consequently, the identification of protein domains is an efficient way to present whether a method can describe the protein structure well. In addition, the connections between the residues are dense within these structural units, which are similar to the connections between communities of the complex networks, expressing the community properties of such network well. To facilitate the understanding of such complex systems, community division was used to analyze these amino acid interaction networks. The purpose of this method is to divide the vertexes of the networks into groups, within which the connections between the vertexes are dense and the connections between which are sparser in the same time [33]. Moreover, a number of the methods based on community have been published in many fields [34–39].

In this study, protein structures were represented by complex networks, in which a vertex is a residue and an edge is an interaction between residues. Here, different cutoff values and strategies used for defining a vertex were tested. For a dataset of 2847 proteins with domain/domains, the identification performance in this study was assessed by accuracy (Acc), which was defined as the proportion of amino acids correctly identified in the certain domain regions of the query sequences according to the information of protein structures in SCOP [40]. For example, suppose the domain regions of the query sequence have 100 amino acids; if 90 of which were correctly identified as belonging to domain regions while the other 10 were misjudged as sequence regions, then the Acc will be 90%. It was observed that when the community division method was based on edge betweenness, the Acc (R_{C_α}) was stable at ~86% when R_{C_α} was around 5.0–7.5 Å, and Acc (R_{C_α}) achieved the highest value of 86.68% when R_{C_α} was 5.0 Å. In addition, when the community division method was based on random walks, the Acc (R_{C_α}) was ~81% when R_{C_α} was around 6.5–7.5 Å, and Acc (R_{C_α}) achieved the highest value of 81.87% when R_{C_α} was 7.0 Å and the step size was 10. The identification performance proved that the optimal cutoff value for constructing the protein structure networks was 5.0 Å (C_α-C_α distances), while the ideal community division method was community structure detection based on edge betweenness in this study. The results suggested that the amino acid interaction networks are an efficient method for describing the structure of proteins, and the different definitions of vertexes and edges do have important effect in this process.

Table 1: The composition of proteins contained in the dataset.

Number of domains	1	2	3	4	5	6	7
Number of proteins	1450	1077	230	66	19	3	2

2. Materials and Methods

2.1. Data Collection and Data Set Construction. The information on domains in proteins in this study were collected from ASTRAL SCOP [40] version 1.75 database. Protein domains in SCOP are grouped into species and hierarchically classified into families, superfamilies, folds, and classes [41]. This database organizes proteins hierarchically according to their families and folds, which is generally considered as the standard for protein structure classification [42]. In order to ensure the nonredundancy of the data, only these proteins with a pairwise sequence identity ≤30% were downloaded, and only those in which the structures were solved by X-ray crystallography with resolution ≤2.5 Å were kept for the clear structure of the proteins. Finally, the remaining 2847 proteins were left for this research. The compositions of the dataset were listed in Table 1.

2.2. Protein Structure Network. Protein structures can be represented as complex networks where amino acids are the nodes and their interactions are the edges [43]. In this study, each protein was considered a small self-governed network system. The structure of proteins was transformed into a complex network by taking amino acid residues as the vertexes and the interactions between the amino acid residues as edges. Various protein structure networks were constructed to investigate the protein structure and the influence of different strategies in building them.

Here, edges are defined in three ways, and from which the optimal cutoff value was finally chosen. Two amino acid residues have a connection if (a) the distance between C_α (defined as R_{C_α}) is 3–10 Å (step size of 0.5 Å, 15 different numerical values in all); (b) the distance between the centers of the side chains (defined as R_{cent}) is 3–10 Å (step size of 0.5 Å, 15 different numerical values in all); (c) the distance between any atoms of the amino acid residues (defined as R_{atm}) is 0–6 Å (step size of 0.5 Å, 13 different numerical values in all). The semidiameters of the atoms were taken into consideration. The amino acid residues interaction networks defined in this study are as shown in Figure 1, 3D structure of which is quite distinct.

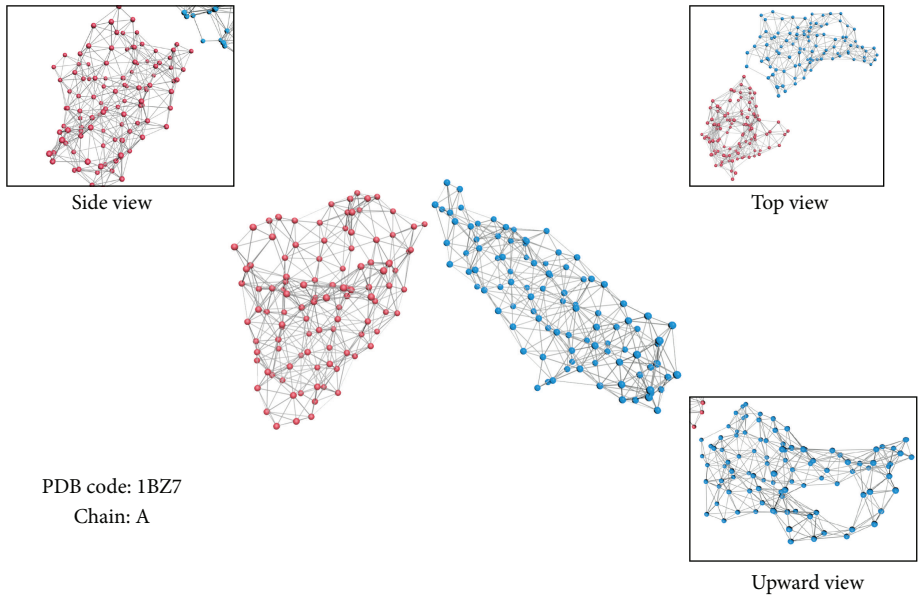

Figure 1: The amino acid residues interaction network. PDB code 1BZ7, chain A. The 3D structure of which is shown above together with its side, top and upward view. Here, the vertex is defined as C_α, and the edge is C_α-C_α distances which is set at 7.5 Å. Each point in the figure represents an amino acid in the chain, which is also the vertex of the network. Ligatures between the vertices are the edge of the network, which illustrate the interaction between the amino acids. For contrasting the figure of community division with this complex network, each vertex is colored based on its identity in SCOP. Here, reddish purple and blue represent different domain regions in this chain.

2.3. Community Division. Tools for network analysis are firmly grounded on the results in graph theory [44], including which network community structure plays an important role in organizing and understanding the complex networks. The network communities were identified as dense groups of the network, whose nodes have a much stronger influence on each other than on the rest of the network [35]. Moreover, the connections between the residues are dense within domains, which express the community properties of such network well. Based on this characteristic, in this study, the community division methods were used to divide the whole sequences into potential domain regions. Two different methods were employed here: community structure detection based on edge betweenness and community structure via short random walks, and between which the more ideal one was finally choosen.

2.3.1. Community Structure Detection Based on Edge Betweenness. Algorithms based on betweenness have been widely applied in various types of networks such as email messages, animal social networks, collaborations of jazz musicians, metabolic networks, and gene networks [33, 45–49]. For more detailed description of this method, refer to papers [45, 50]. The principle of the community structure detection based on edge betweenness is that it seems that all the shortest paths from one module to another must traverse through the edges connecting separate modules, which have high edge betweenness in that case.

As a result, this algorithm is performed by calculating the edge betweenness of the graph and removing the edge with the highest edge betweenness score gradually in order to obtain a hierarchical map. This rooted tree is the dendrogram of the graph, the leaves are the individual vertices, and the roots represent the whole graph. Finally, a numeric matrix is constructed using this algorithm.

2.3.2. Community Structure via Short Random Walks. Algorithms based on random walks have been applied in various researches of networks [50, 51]. This algorithm tries to find densely connected subgraphs which are also known as communities in a graph via short random walks. The principle of this algorithm is that short random walks are likely to stay in the same community. It takes every single node as an independent community at first, then those of which tally with certain rules were incorporated together step by step. It introduces r as a distance between the vertices, which shall be small if the two vertices are in the same community and large if they are not.

3. Results and Discussion

3.1. Community Division Based on Edge Betweenness. In this section, community division method based on edge betweenness was applied on complex networks, and the effect of different cutoff values of edges for constructing complex networks was analyzed. Then, an optimized cutoff value was identified. The flowchart of these two steps, amino acid interaction network together with community division methods, is shown in Figure 2.

For the fairness of the contrast, all complex networks constructed by different cutoff values were analyzed by community division method, which insures the most optimal

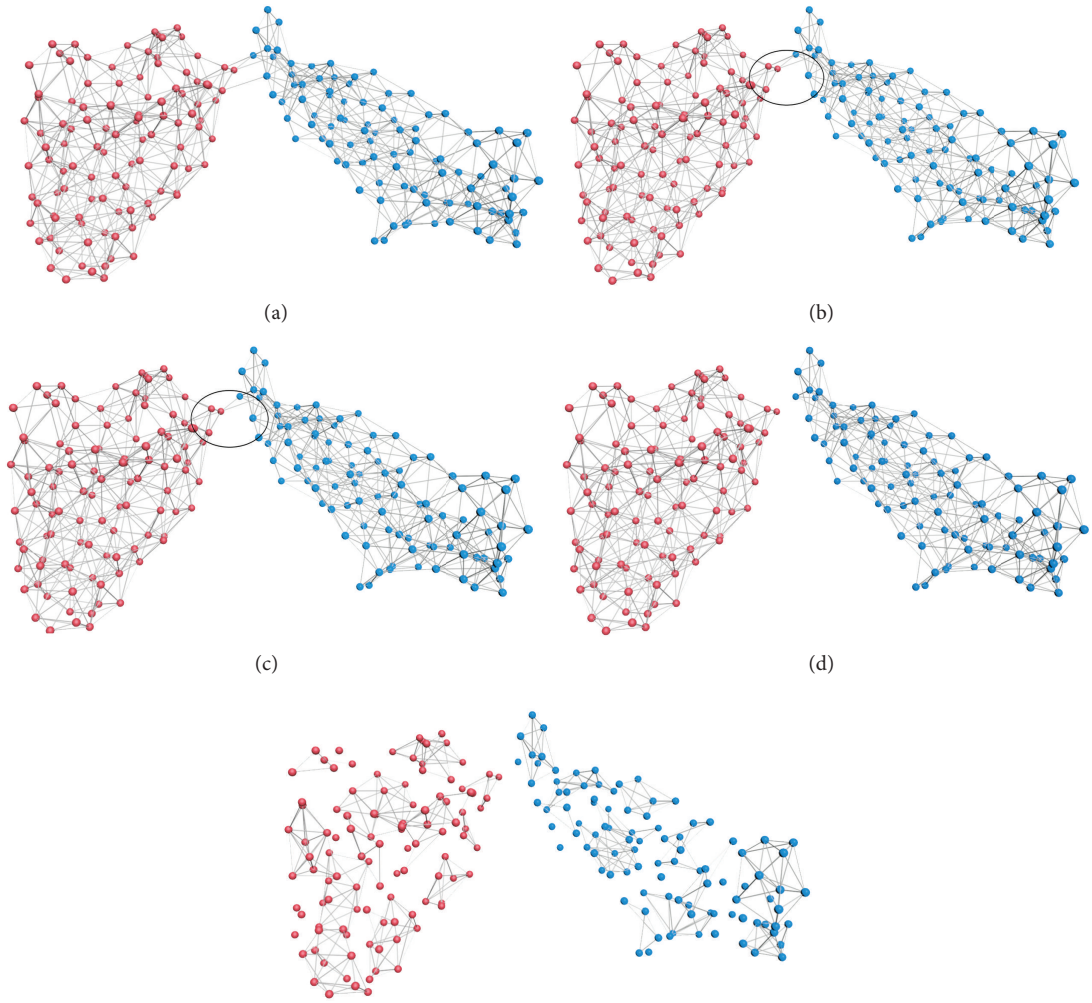

Figure 2: The flowchart of the amino acid interaction network together with community division method. PDB code 1BZ7, chain A. Each point in the figure represents an amino acid in the chain, which is also the vertex of the network. Ligatures between the vertices are the edge of the network, which illustrate the interaction between the amino acids. Here, the reddish purple and blue represent different domain regions in this chain based on the identity in SCOP. Firstly, an amino acid complex network was constructed with the vertex defined as C_α, and the edge as C_α-C_α distance which was set at 7.5 Å, as shown in (a). Secondly, community division was based on edge betweenness, and the first edge with the highest edge betweenness score was removed, as shown in (b). Thirdly, more edges were removed based on the algorithm, and (c) shows that three edges were removed. Fourthly, the community division was finished when the correct number of edges was removed, as shown in (d); two different domains have been clearly separated, and five edges were removed for this protein. Finally, if the community division is taken continually, more communities will be found in the complex network. (e) shows the result of community division for chain A of protein 1BZ7 after removing 500 edges in this complex network, and many more communities illustrate the wrong results according to the identity in SCOP.

results. In order to obtain the best prediction performance, different cutoff values were evaluated based on multidomain proteins. 15 different values (3–10 Å) of the R_{C_α} and the R_{cent} (step size of 0.5 Å) were optimized, respectively, and so were other 13 different distance values (0–6 Å) of R_{atom} (step size of 0.5 Å).

First, threshold of 7 Å, which has been reported to be an important distance parameter because all contacts are complete and legitimate (not occluded) at this distance [23], was analyzed. The results were obtained after the community division. The identification performance in this study was assessed by accuracy, which was defined as the proportion of amino acids correctly identified in the certain domain regions of the query sequences. When the R_{C_α} and the R_{cent} were 7 Å, respectively, the results are 86.21% and 85.16%, respectively.

More cutoff values were tested via different strategies of vertex. First, the average accuracies for all the proteins defined by R_{C_α} were listed in Table 2. The results indicated

Table 2: The accuracies of all proteins defined by R_{C_α} based on edge betweenness.

Threshold	Accuracy
3 Å	2.15
3.5 Å	2.17
4 Å	78.96
4.5 Å	83.42
5 Å	86.68
5.5 Å	86.45
6 Å	85.54
6.5 Å	85.76
7 Å	86.21
7.5 Å	85.92
8 Å	85.21
8.5 Å	84.75
9 Å	84.28
9.5 Å	83.71
10 Å	83.86

Table 3: The accuracies of all proteins defined by R_{cent} based on edge betweenness.

Threshold	Accuracy
3 Å	2.14
3.5 Å	2.59
4 Å	3.79
4.5 Å	7.42
5 Å	33.99
5.5 Å	78.87
6 Å	84.53
6.5 Å	85.04
7 Å	85.16
7.5 Å	85.52
8 Å	84.89
8.5 Å	84.48
9 Å	83.83
9.5 Å	83.56
10 Å	83.40

that when the method was based on the edge betweenness, Acc (R_{C_α}) achieved the highest 86.68% when R_{C_α} was 5.0 Å. When R_{C_α} was around 5.0–7.5 Å, the accuracies were around 86%, and the bias of the numerical values in this area was small (~1%). This illustrated that the cutoff values in this area reflected protein structure well. Second, the average accuracies for all the proteins defined by R_{cent} were listed in Table 3. The results indicated that Acc (R_{cent}) achieved the highest 85.52% when R_{cent} was 7.5 Å. When R_{cent} was around 6.5–8.0 Å, Acc (R_{cent}) showed relatively ideal values around 85%, which illustrated that the cutoff values in this area reflected protein structure well. However, the bias of the numerical values was evident for all the numerical values of R_{cent}. Acc (R_{cent}) were lower than 10% when R_{cent} was around 3.0–4.5 Å, which were generated by the otherness

Table 4: The accuracies of all proteins defined by R_{atom} based on edge betweenness.

Threshold	Accuracy
0 Å	85.06
0.5 Å	85.36
1.0 Å	85.58
1.5 Å	85.59
2 Å	85.06
2.5 Å	84.39
3 Å	83.73
3.5 Å	83.50
4 Å	83.95
4.5 Å	83.93
5 Å	83.51
5.5 Å	83.45
6 Å	83.31

of the size of side chains. Third, the average accuracies for all the proteins defined by R_{atom} were listed in Table 4. The results indicated that when the distance between any atoms of the amino acid residues defined as R_{atom} was taken into consideration, the superiority of the diversity of the volume of atoms should also be taken into consideration. Acc(R_{atom}) achieved the highest value of 85.59% when R_{atom} was 1.5Å. When R_{atom} was around 0.0–2.0 Å, Acc (R_{atom}) showed relatively ideal values around 85%, and the bias of the numerical values in this area was small (~0.6%). When the cutoff values were bigger than 2.0 Å, Acc (R_{atom}) decreased monotonically as R_{atom} increased. That is, overlarge R_{atom} will lead to the incorrect identification of the interactions among amino acids, which will distort the actual protein structure.

It was observed that when the community division method was based on edge betweenness, the Acc (R_{C_α}) was stable at ~86%, which illustrated that the network characterization of protein structure would not be limited by its type. Furthermore, Acc (R_{cent}) was ~1% lower than that of Acc (R_{C_α}), which was generated by the cutoff value. That is, the side chains of the amino acids have a certain space volume, and a big cutoff value signifies the space overlap of the atoms from different amino acids, which is obviously inappropriate for protein structure. In conclusion, Acc (R_{cent}) was lower than Acc (R_{C_α}) and Acc (R_{atom}), which illustrated that the space specificity of the side chains of amino acids affects the construction of the amino acids complex networks. It was observed that the highest accuracy obtained was 86.68% (R_{C_α} = 5.0 Å). That is, the optimal cutoff value was 5.0 Å (C_α-C_α distances) when the ideal community division method was based on edge betweenness.

3.2. Community Division Based on Random Walks. In this section, the community division method based on random walks was analyzed. The same cutoff values were evaluated here based on multidomain proteins, that is, 15 different numerical values (3–10 Å) of the R_{C_α} and the R_{cent} (step size of 0.5 Å) and other 13 different numerical values (0–6 Å) of

Table 5: Acc (R_{C_α}) and Acc (R_{cent}) of all proteins based on random walks under 7 Å of different step sizes.

Step size	3	4	5	6	7	8	9	10
Acc (R_{C_α})	77.37	78.56	79.84	80.21	80.93	81.23	81.43	81.93
Acc (R_{cent})	76.39	77.62	78.56	79.12	79.64	80.05	80.13	80.70

Table 6: The accuracies of all proteins defined by R_{C_α} based on random walks.

Threshold	Accuracy
3 Å	0
3.5 Å	0
4 Å	67.14
4.5 Å	69.65
5 Å	73.84
5.5 Å	79.87
6 Å	80.39
6.5 Å	81.09
7 Å	81.93
7.5 Å	81.85
8 Å	80.97
8.5 Å	80.48
9 Å	80.46
9.5 Å	79.95
10 Å	79.71

R_{atom} (using a step size of 0.5 Å). In addition, the step sizes of the community division based on random walks were also optimized here.

First, threshold of 7 Å [23] was analyzed for all the proteins. When the R_{C_α} and the R_{cent} were 7 Å, respectively, the results are listed in Table 5.

It was observed that when the community division method was based on random walks under the threshold of 7 Å via different step sizes, the highest Acc (R_{C_α}) and Acc (R_{cent}) were 81.93% and 80.70%, respectively. The numeric values of them all were ~4% lower than that for edge betweenness, which was generated by the method itself. That is, the algorithm based on the random walks attempted to find a given length called step size, which is obviously inappropriate for domains of different sizes. In large domains, a short length will not project all the amino acids in the same community.

More cutoff values were tested via different strategies of vertex. First, the average accuracies for all the proteins defined by R_{C_α} were listed in Table 6. The results indicated that Acc (R_{C_α}) achieved the highest 81.87% when R_{C_α} was 7.0 Å and the step size was 10. When R_{C_α} was around 6.5–7.5 Å, the accuracies were around 81%, and the bias of the numerical values in this area was small (~1%). This illustrated that the cutoff values in this area reflected protein structure well. However, the numeric of Acc (R_{C_α}) was ~5% lower than that for edge betweenness. Second, the average accuracies for all the proteins defined by R_{cent} were listed in Table 7. The results indicated that Acc (R_{cent}) achieved the highest value of

Table 7: The accuracies of all proteins defined by R_{cent} based on random walks.

Threshold	Accuracy
3 Å	0
3.5 Å	0
4 Å	0
4.5 Å	0
5 Å	0
5.5 Å	5.05
6 Å	59.20
6.5 Å	78.34
7 Å	80.63
7.5 Å	80.63
8 Å	80.77
8.5 Å	80.20
9 Å	79.60
9.5 Å	79.64
10 Å	79.41

Table 8: The accuracies of all proteins defined by R_{atom} based on random walks.

Threshold	Accuracy
0 Å	80.39
0.5 Å	80.58
1.0 Å	80.82
1.5 Å	80.70
2 Å	80.79
2.5 Å	80.08
3 Å	79.55
3.5 Å	79.35
4 Å	79.24
4.5 Å	78.98
5 Å	78.68
5.5 Å	78.36
6 Å	77.49

80.77% when R_{cent} was 8.0 Å and the step size was 10. When R_{cent} was around 7.0–8.5 Å, Acc (R_{cent}) showed relatively ideal values around 80%, which illustrated that the cutoff values in this area reflected protein structure well. However, the bias of the numerical values was evident for all the numerical values of R_{cent}, which were generated by the otherness of the side chains. The numeric of Acc (R_{cent}) was ~5% lower than that for edge betweenness, and Acc (R_{cent}) was as low as 0% when R_{C_α} was around 3.0–5 Å, which may be produced by the looseness of the complex networks constructed under these thresholds. Third, the average accuracifes for all the proteins defined by R_{atom} were listed in Table 8. The results indicated that when the distance between any atoms of the amino acid residues defined as R_{atom} was taken into consideration, the superiority of the diversity of the volume of atoms should also be taken into consideration. Acc (R_{atom}) achieved the highest value of 80.82% when R_{atom} was 1.0 Å and the step

Table 9: The optimal accuracies of each dataset based on edge betweenness.

Dataset	1	2	3	4	5	6	7	8
R_{C_α}	7.00 Å	5.50 Å	5.50 Å	5.00 Å	5.50 Å	5.00 Å	5.50 Å	5.50 Å
Accuracy	84.67	89.08	87.07	86.52	87.35	87.26	86.95	86.50
R_{cent}	6.50 Å	7.50 Å	7.50 Å	7.50 Å	7.50 Å	7.50 Å	7.50 Å	7.50 Å
Accuracy	82.51	86.93	86.50	85.74	86.17	86.58	85.85	85.49
R_{atom}	1.00 Å	1.00 Å	0.50 Å	1.00 Å	1.50 Å	1.00 Å	1.00 Å	1.00 Å
Accuracy	82.89	87.54	86.24	86.13	86.94	86.61	85.61	85.80

Table 10: The optimal accuracies of each dataset based on random walks.

Dataset	1	2	3	4	5	6	7	8
R_{C_α}	6.00 Å	7.50 Å	7.50 Å	7.50 Å	7.50 Å	7.00 Å	7.50 Å	7.00 Å
Step size	10	10	10	10	10	10	10	10
Accuracy	75.34	85.00	82.46	81.61	83.20	83.39	82.25	81.93
R_{cent}	7.00 Å	7.00 Å	8.00 Å	8.00 Å	8.00 Å	8.00 Å	7.50 Å	7.00 Å
Step size	10	10	10	10	9	10	10	10
Accuracy	74.62	84.95	80.97	80.89	81.84	82.67	80.61	80.79
R_{atom}	0.50 Å	1.50 Å	0.50 Å	1.00 Å	1.50 Å	1.00 Å	1.00 Å	1.00 Å
Step size	10	10	10	9	10	10	10	10
Accuracy	74.85	84.66	81.20	81.11	82.36	82.97	81.45	80.95

size was 10. When R_{atom} was around 0.0–2.5 Å, Acc (R_{atom}) showed relatively ideal values around 80%, and the bias of the numerical values in this area was small (~1%). However, the numeric of Acc (R_{atom}) was 5% lower than that for edge betweenness.

In conclusion, Acc (R_{cent}) was lower than Acc (R_{C_α}) and Acc (R_{atom}). It was observed that when the community division method was based on random walks, the numeric of the accuracy was lower than that based on edge betweenness all the while, which indicated that the ideal community division method for this research was community structure detection based on edge betweenness. Moreover, the value of Acc (R_{cent}) was the worst via both the two community division methods all along. Similar results were obtained in the study of side chain contact models; three models were compared and the isotropic sphere side chain (ISS) model was the worst in accuracy. They proved that the model which took the spatially anisotropic nature of the side chain into consideration would eliminate about 95% of the incorrectly counted contact pairs in the ISS model [26]. However, this kind of practical models do have less moderate computational cost than the popular representation model such as the use of C_α atom, which is proved to be effective for the kind of the data in this study.

3.3. The Stability Analysis of the Method. To verify the stability of the method, 8 datasets were constructed based on multidomain proteins. The first dataset was composed of 100 proteins, and every other dataset contained 100 proteins more than the previous one. That is, the 8th dataset contained 800 proteins.

The same operations were taken based on these 8 datasets. Different numerical values of R_{C_α} (3–10 Å), R_{cent} (3–10 Å), and R_{atom} (0–6 Å) were optimized based on two community division methods. The highest accuracies for each dataset were listed in Tables 9 and 10.

It was observed that when the community division method was based on edge betweenness, Acc (R_{C_α}) for each database got the highest results around ~86%–89% when R_{C_α} was ~5.00–5.50 Å, which were quite close to the result 86.68% when R_{C_α} was 5.00 Å. However, results for database one was a little bit different, 84.67% when R_{C_α} was 7.00 Å, which may be generated by the lack of statistically significant result in the small amount of the proteins. Acc (R_{cent}) for each database got the highest results around ~85%-86% when R_{cent} was 7.50 Å, which were quite close to the result 85.52% when R_{cent} was 7.50 Å. However, results for database one was a little bit different, 82.51% when R_{cent} was 6.50 Å, which may be generated by the lack of statistically significant result in the small amount of the proteins. Acc (R_{atom}) for each database got the highest results around ~82%–87% when R_{atom} was ~0.50–1.50 Å, which were quite close to the result 85.59% when R_{C_α} was 1.50 Å.

When the community division method was based on random walks, Acc (R_{C_α}) for each database got the highest results around ~81%–85% when R_{C_α} was ~7.00–7.50 Å and the step size was 10, which were quite close to the result 81.87% when R_{C_α} was 7.0 Å and the step size was 10. Acc (R_{cent}) for each database got the highest results around ~80%–84% when R_{cent} was 7.00–8.00 Å, which were quite close to the result 80.77% when R_{cent} was 8.0 Å and the step size was 10. Acc (R_{atom}) for each database got the highest results around ~80%–84% when R_{atom} was ~0.50–1.50 Å and the step size was 10, which were quite close to the result 80.82% when R_{C_α} was 1.00 Å and the step size was 10. However, results for database one was a little bit different under these three

conditions, which may be generated by the lack of statistically significant result in the small amount of the proteins.

It is observed from the results that the complex networks together with the community division methods constructed in this study were stable, which proved the creditability of the research. On the other hand, it was observed that when the community division method was based on edge betweenness, the Acc (R_{C_α}) was stable at ~86% when R_{C_α} was around 5.0–7.5 Å, and the optimal cutoff value for constructing the protein structure networks was 5.0 Å (C_α-C_α distances) in this study.

4. Conclusion

The main objective of this study is to explore the contribution of complex network together with its different definitions of vertexes and edges to describing the structure of proteins. When applying our method on a dataset of 2847 proteins with domain/domains, it was observed that when the community division method was based on random walks, the numeric of the accuracy was lower than that based on edge betweenness all the while, which indicated that the ideal community division method for this research was community structure detection based on edge betweenness. When the community division method was based on edge betweenness, the Acc (R_{C_α}) was stable at ~86% when R_{C_α} was around 5.0–7.5 Å, and Acc (R_{C_α}) achieved the highest value of 86.68% when R_{C_α} was 5.0 Å. The identification performance proved that the optimal cutoff value for constructing the protein structure networks was 5.0 Å (C_α-C_α distances), while the ideal community division method was community structure detection based on edge betweenness in this study. The results suggested that the amino acid interaction networks are an efficient method for describing the structure of proteins, and the different definitions of vertexes and edges do have important effect in this process. Distance should be taken into consideration to prevent unnecessary deviation. Moreover, the optimized network model could be further applied in future study for the number and position of protein domain prediction.

Acknowledgments

The authors would like to thank the anonymous reviewers for their patient review and constructive suggestions. This study was supported by the Natural Science Foundation of China (21175095, 20972103).

References

[1] R. C. Penner, M. Knudsen, C. Wiuf, and J. E. Andersen, "An algebro-topological description of protein domain structure," *PLoS ONE*, vol. 6, no. 5, article e19670, Article ID e19670, 2011.

[2] A. L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[3] O. Magger, Y. Y. Waldman, E. Ruppin, and R. Sharan :, "Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks," *PLoS Computational Biology*, vol. 8, no. 9, article e1002690, 2012.

[4] S. Karni, H. Soreq, and R. Sharan, "A network-based method for predicting disease-causing genes," *Journal of Computational Biology*, vol. 16, no. 2, pp. 181–189, 2009.

[5] F. Cheng, C. Liu, J. Jiang et al., "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Computational Biology*, vol. 8, no. 5, article e1002503, 2012.

[6] P. Csermely, V. Ágoston, and S. Pongor, "The efficiency of multi-target drugs: the network approach might help drug design," *Trends in Pharmacological Sciences*, vol. 26, no. 4, pp. 178–182, 2005.

[7] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, 2001.

[8] R. Albert and A. L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.

[9] M. Vendruscolo, N. V. Dokholyan, E. Paci, and M. Karplus, "Small-world view of the amino acids that play a key role in protein folding," *Physical Review E*, vol. 65, no. 6, article 061910, Article ID 061910, 4 pages, 2002.

[10] N. V. Dokholyan, L. Li, F. Ding, and E. I. Shakhnovich, "Topological determinants of protein folding," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 13, pp. 8637–8641, 2002.

[11] K. V. Brinda and S. Vishveshwara, "A network representation of protein structures: implications for protein stability," *Biophysical Journal*, vol. 89, no. 6, pp. 4159–4170, 2005.

[12] B. Thibert, D. E. Bredesen, and G. del Rio, "Improved prediction of critical residues for protein function based on network and phylogenetic analyses," *BMC Bioinformatics*, vol. 6, article 213, 2005.

[13] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular systems biology*, vol. 3, article 88, 2007.

[14] C. Böde, I. A. Kovács, M. S. Szalay, R. Palotai, T. Korcsmáros, and P. Csermely, "Network analysis of protein dynamics," *FEBS Letters*, vol. 581, no. 15, pp. 2776–2782, 2007.

[15] R. Konrat, "The protein meta-structure: a novel concept for chemical and molecular biology," *Cellular and Molecular Life Sciences*, vol. 66, no. 22, pp. 3625–3639, 2009.

[16] G. Amitai, A. Shemesh, E. Sitbon et al., "Network analysis of protein structures identifies functional residues," *Journal of Molecular Biology*, vol. 344, no. 4, pp. 1135–1146, 2004.

[17] N. T. Doncheva, K. Klein, F. S. Domingues, and M. Albrecht, "Analyzing and visualizing residue networks of protein structures," *Trends in Biochemical Sciences*, vol. 36, no. 4, pp. 179–182, 2011.

[18] A. J. M. Martin, M. Vidotto, F. Boscariol, T. Di Domenico, I. Walsh, and S. C. E. Tosatto, "RING: networking interacting residues, evolutionary information and energetics in protein structures," *Bioinformatics*, vol. 27, no. 14, pp. 2003–2005, 2011.

[19] Y. Assenov, F. Ramírez, S. E. S. E. Schelhorn, T. Lengauer, and M. Albrecht, "Computing topological parameters of biological networks," *Bioinformatics*, vol. 24, no. 2, pp. 282–284, 2008.

[20] N. T. Doncheva, Y. Assenov, F. S. Domingues, and M. Albrecht, "Topological analysis and interactive visualization of biological networks and protein structures," *Nature Protocols*, vol. 7, no. 4, pp. 670–685, 2012.

[21] A. R. Atilgan, P. Akan, and C. Baysal, "Small-world communication of residues and significance for protein dynamics," *Biophysical Journal*, vol. 86, no. 1, pp. 85–91, 2004.

[22] G. Bagler and S. Sinha, "Network properties of protein structures," *Physica A*, vol. 346, no. 1-2, pp. 27–33, 2005.

[23] C. H. Da Silveira, D. E. V. Pires, R. C. Minardi et al., "Protein cutoff scanning: a comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins," *Proteins*, vol. 74, no. 3, pp. 727–743, 2009.

[24] E. Estrada, "Universality in protein residue networks," *Biophysical Journal*, vol. 98, no. 5, pp. 890–900, 2010.

[25] L. H. Greene and V. A. Higman, "Uncovering network systems within protein structures," *Journal of Molecular Biology*, vol. 334, no. 4, pp. 781–791, 2003.

[26] W. Sun and J. He, "From isotropic to anisotropic side chain representations: comparison of three models for residue contact estimation," *PLoS ONE*, vol. 6, no. 4, Article ID e19238, 2011.

[27] J. T. Guo, D. Xu, D. Kim, and Y. Xu, "Improving the performance of DomainParser for structural domain partition using neural network," *Nucleic Acids Research*, vol. 31, no. 3, pp. 944–952, 2003.

[28] Y. Xu, D. Xu, and H. N. Gabow, "Protein domain decomposition using a graph-theoretic approach," *Bioinformatics*, vol. 16, no. 12, pp. 1091–1104, 2000.

[29] J. E. Gewehr and R. Zimmer, "SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles," *Bioinformatics*, vol. 22, no. 2, pp. 181–187, 2006.

[30] J. Cheng, M. J. Sweredoski, and P. Baldi, "DOMpro: protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks," *Data Mining and Knowledge Discovery*, vol. 13, no. 1, pp. 1–10, 2006.

[31] J. S. Richardson, "The anatomy and taxonomy of protein structure," *Advances in Protein Chemistry*, vol. 34, pp. 167–339, 1981.

[32] D. B. Wetlaufer, "Nucleation, rapid folding, and globular intrachain regions in proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 70, no. 3, pp. 697–701, 1973.

[33] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.

[34] M. Szalay-Beko, R. Palotai, B. Szappanos, I. A. Kovacs, B. Papp, and P. Csermely :, "ModuLand plug-in for Cytoscape: determination of hierarchical layers of overlapping network modules and community centrality," *Bioinformatics*, vol. 28, no. 16, pp. 2202–2204, 2012.

[35] I. A. Kovács, R. Palotai, M. S. Szalay, and P. Csermely, "Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics," *PLoS ONE*, vol. 5, no. 9, article e12528, Article ID e12528, pp. 1–14, 2010.

[36] Y. Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, 2010.

[37] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75–174, 2010.

[38] A. Delmotte, E. W. Tate, S. N. Yaliraki, and M. Barahona, "Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin-myosin light chain interaction," *Physical Biology*, vol. 8, no. 5, article 055010, 2011.

[39] J. C. Delvenne, S. N. Yaliraki, and M. Barahon, "Stability of graph communities across time scales," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 29, pp. 12755–12760, 2010.

[40] S. E. Brenner, P. Koehl, and M. Levitt, "The ASTRAL compendium for protein structure and sequence analysis," *Nucleic Acids Research*, vol. 28, no. 1, pp. 254–256, 2000.

[41] L. Lo Conte, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin, "SCOP database in 2002: refinements accomodate structural genomics," *Nucleic Acids Research*, vol. 30, no. 1, pp. 264–267, 2002.

[42] R. Day, D. A. C. Beck, R. S. Armen, and V. Daggett, "A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary," *Protein Science*, vol. 12, no. 10, pp. 2150–2160, 2003.

[43] S. Lifson and C. Sander, "Antiparallel and parallel β-Strands differ in amino acid residue preferences," *Nature*, vol. 282, no. 5734, pp. 109–111, 1979.

[44] L. D. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, "Characterization of complex networks: a survey of measurements," *Advances in Physics*, vol. 56, no. 1, pp. 167–242, 2007.

[45] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, Article ID 026113, p. 1, 2004.

[46] D. M. Wilkinson and B. A. Huberman, "A method for finding communities of related genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 1, pp. 5241–5248, 2004.

[47] P. Holme, M. Huss, and H. Jeong, "Subnetwork hierarchies of biochemical pathways," *Bioinformatics*, vol. 19, no. 4, pp. 532–538, 2003.

[48] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, "E-Mail as spectroscopy: automated discovery of community structure within organizations," *Information Society*, vol. 21, no. 2, pp. 133–153, 2005.

[49] P. M. Gleiser and L. Danon, "Community structure in jazz," *Advances in Complex Systems*, vol. 6, no. 4, pp. 565–573, 2003.

[50] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *Proceedings of the Computer and Information Sicences (ISCIS '05)*, vol. 3733, pp. 284–293, 2005.

[51] H. J. Zhou and R. Lipowsky, "Network brownian motion: a new method to measure vertex-vertex proximity and to identify communities and subcommunities," in *Proceedings of the Computational Science (ICCS '04)*, vol. 3038, Part 3, pp. 1062–1069, 2004.