Residue network in protein native structure belongs to the universality class of three dimensional critical percolation cluster

Hidetoshi Morita and Mitsunori Takan
Faculty of Science and Engineering, Waseda University, Tokyo 169-8555, Japan
(Dated: September 28, 2008)

A single protein molecule is regarded as a contact network of amino-acid residues. Some studies have indicated that this network is a small world network (SWN), while other results have implied that this is a fractal network (FN). However, SWN and FN are essentially different in the dependence of the shortest path length on the number of nodes. In this paper, we investigate this dependence in the residue contact networks of proteins in native structures, and show that the networks are not SWN but FN. FN is generally characterized by several dimensions. Among them, we focus on three dimensions; the network topological dimension $D_c$, the fractal dimension $D_f$, and the spectral dimension $D_s$. We find that proteins universally yield $D_c \approx 1.9$, $D_f \approx 2.5$ and $D_s \approx 1.3$. These values are in surprisingly good coincidence with those in three dimensional critical percolation cluster. Hence the residue contact networks in the protein native structures belong to the universality class of three dimensional percolation cluster. The criticality is relevant to the ambivalent nature of the protein native structures, i.e., the coexistence of stability and instability, both of which are necessary for a protein to function as a molecular machine or an allosteric enzyme.

PACS numbers: 87.15.-V, 64.60.-i, 89.75.He, 05.45.Df

Introduction Proteins are one-dimensional chains of amino-acid residues embedded in three dimensional ($D = 3$; 3D) Euclidean space. The residues neighboring in the Euclidean space are in contact with each other. Thus we can regard a protein molecule as a contact network of amino-acid residues [1,2]. This network viewpoint is complementary to the energy landscape picture [3] in understanding the general properties of proteins. We hereafter consider this network within single protein molecules in their native structures, in particular focusing on its universality among proteins.

Some recent studies [4,5,6,7,8] have applied the latest network theory to the residue network, by regarding the amino-acid residues and their contacts as nodes and edges, respectively. The important quantities to characterize the network are the clustering coefficient $C$ and the shortest path length $L$. Those studies have demonstrated that in the residue networks $C$ is larger than the random networks [10] while $L$ is smaller than the normal lattice. This indicates that the residue network is a small world network (SWN) [11].

On the other hand, the spacial profile of residues within single protein molecules has long been studied with the use of authentic methods of material science. Earlier spectroscopic studies [12] have shown anomalous density of states. These results, accompanied with theoretical studies [13], have suggested that the protein structures possess the property of fractal lattice. The fractality within single proteins has also been supported numerically through the density of normal modes [14,15,16] and the spacial mass distribution [17]. This implies that the residue network that we are interested in is a fractal network (FN).

From the general viewpoint of the network theory, however, there lies a dichotomy between SWN and FN [18]. The clustering coefficient $C$ cannot discriminate between SWN and FN, since in both the networks $C$ have a larger value than the random networks. In contrast, the dependence of the shortest path length $L$ on the number of nodes $N$ is essentially different between SWN and FN; $L$ depends on $N$ logarithmically and algebraically, respectively. By exploiting the $N$-dependency of $L$, we can differentiate SWN and FN, in principle.

In proteins, nevertheless, it is practically difficult to clearly distinguish between these two $N$-dependence. This is because the size of proteins does not distribute widely enough to cover sufficient decades. The same data sets can be read as a straight line both in log-log (SWN) and semi-log (FN) plot.

To overcome this difficulty, here we introduce a more sophisticated method. Instead of the $N$-$L$ plot among various sized proteins, we investigate an equivalent within single protein molecules; we calculate the number of nodes $n_l$ that can be reached until $l$ path steps. Then, by overdrawing the $n_l$-$l$ plot for various sized proteins, we obtain a universal curve, as well as the deviation from it due to finite size effect. Thus we can discuss an asymptotic behavior in the large $N$ limit. We thereby find that network in protein native structures is FN, not SWN. This is the first result of this letter.

We then obtain the three characteristic dimensions of fractal residue network; the network topological dimension $D_c$, the fractal dimension $D_f$, and the spectral dimension $D_s$. The values of them are universal among single-chain proteins. Furthermore, these three values surprisingly coincide with those of the 3D critical percolation cluster. Namely, proteins belongs to the universality class of 3D critical percolation cluster. This is the second and the most highlighted result of this letter.
Dichotomy between FN and SWN [18].

ships (2) and (3) are essentially different, leading to the relationship is [18],

\[ D \sim l/l_0, \quad n_l \sim \exp(l/l_0). \]

for a positive constant \( l_0 \). Note again that the relationships [20] and [21] are essentially different, leading to the dichotomy between FN and SWN [18].

FIG. 1 shows the relationship between \( n_l \) and \( l \); the same data sets are plotted in (a) log-log and (b) semi-log scales. We present the data for five representative proteins of different size: ribonuclease T1 (PDB ID=9RNT, 104 amino acids (a.a.)), cutinase (1CUS, 200 a.a.), green fluorescent protein (1EMA, 236 a.a.), actin, (1J6Z, 375 a.a.), and subfragment 1 of myosin (1SR6, 1152 a.a.). Obviously the data obeys the power-law scaling better than the exponential dependence. This is also supported by considering the finite size effect as follows. In (a), the range where the data follow the power-law scaling tends to extend as the number of nodes \( N \) increases. This suggests the existence of an asymptotic universal line in the limit of \( N \to \infty \). In (b), on the contrary, we cannot see such an asymptotic tendency. Thus we conclude that proteins universally obey the power-law scaling (2) with \( D_c \approx 1.9 \). Hence the networks in protein native structures are FN, not SWN.

In much larger proteins, \( D_c \) often gives a bit larger value than 1.9, or even the scaling itself is smeared. This is because the larger proteins are usually not single-domain nor single-chain but multi-domain or multi-chain proteins. Even in such proteins, however, each single-domain or single-chain component still yields the same scaling law with the same dimension \( D_c \approx 1.9 \) [20].

One plausible reason why the network is not SWN but FN is that the residues are spatially restricted in the 3D Euclidean space. Indeed, it has been suggested that networks with spatial (geographical) restriction tend to be rather regular (including fractal) network than SWN [18].

Finally, as we have mentioned, the fractal dimension of the network is determined from the spacial distribution of nodes. Here we again employ the method within single proteins, differently from the previous studies [17], in order to discuss the asymptotic behavior in the limit \( N \to \infty \). Let \( n^{(i)}(d) \) be the number of nodes the distance of which from the node \( i \) is less than \( d \); \( n^{(i)}(d) = \sum_j \Theta(d - d_{ij}) \). Since we are interested in the overall network property of a protein, we consider its average, \( n(d) = \sum_i n^{(i)}(d)/N \), that is,

\[ n(d) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \Theta(d - d_{ij}). \]

Note that this is nothing but the correlation integral introduced by Grassberger and Procaccia [22], although this is not normalized in order to consider the finite size effect. As \( d \) becomes larger, \( n(d) \) monotonically increases and finally saturates at \( N \). In the \( D \)-dimensional normal lattice, \( n(d) \sim d^D \). Similarly, if the spacial distribution of nodes is fractal,

\[ n(d) \sim d^{D'}, \]
where $D_f$ is referred to as the fractal dimension.

FIG. 2 shows $n(d)$ versus $d$ in log-log scale, for the same proteins as FIG. 1. The relationship follows power-law scaling. Similarly to the case in the network topological dimension, this is supported by considering the finite size effect; the power-law scaling range tends to extend as the number of nodes increases, suggesting the existence of an asymptotic universal line in the limit $N \to \infty$. Thus we conclude that proteins universally follows the power-law scaling $D_f \approx 2.5$, which is consistent with the previous studies [17].

Spectral dimension. The spectral dimension is determined from the density of normal modes (DNM). According to the Debye theory, DNM in $D$-dimensional normal lattice is $\rho(\omega) \sim \omega^{D-1}$. Similarly, DNM in FN obeys,

$$\rho(\omega) \sim \omega_{D_s-1},$$

where $D_s$ is referred to as the spectral dimension [22].

DNM is, in general, obtained experimentally by spectroscopies and numerically by normal mode analysis (NMA). To be relevant to experiments, we conduct NMA in the all atom model, not in a coarse grained model. Then, by focusing on the frequency region corresponding to the residue-residue interaction, we consider the spectral dimension of the residue network. We do so because for NMA it is necessary to take the interaction strengths precisely into account. In the all atom model, the interaction strengths are quite reliable, since it is basically obtained from quantum chemical calculations. In a coarse grained model, in contrast, the interaction strengths are introduced rather arbitrary. It is true that the coarse grained models well reproduces the overall fluctuation of the protein native structure [2]. This is, however, largely due to the fact that only a limited number of lowest frequency normal modes (or largest amplitude principal components) dominate the fluctuation. There is no guarantee that they also reproduce DNM for decades. Indeed, it has been reported that there is an essential difference in DNM between the all atom model and the coarse grained model with identical interaction strengths [24]. Instead, here we coarse grain DNM itself, by truncating the higher frequency region. We perform NMA by using the program NMODE implemented in the AMBER software [25], with AMBER force field (perm99) and implicit water (Generalized Born) model. Before NMA, energy minimization is executed with Newton-Raphson and conjugate gradient method, so that the norm of the force is less than the order of $10^{-12}$ kcal mol$^{-1}$ Å$^{-1}$.

We have obtained DNM for several proteins, and FIG. 3 shows typical results; these are essentially similar to one of the previous numerical studies [16]. There exist two shoulders at around 10 and 100 cm$^{-1}$, which are denoted respectively by $\omega_{FS}$ and $\omega_{GL}$. The frequency higher than $\omega_{GL}$ corresponds to local motions, due to covalent-bond stretching and angle bending motions. The frequency lower than $\omega_{GL}$, in contrast, corresponds to global motions due to residue-residue interactions, which we are now interested in. In the latter region, DNM obeys the power-law scaling [17] with $D_s \approx 1.3$. At around $\omega_{FS}$, the dimension changes from 1.3 to 3.0. This is due to the finite size effect; through a long wave-length probe, the protein is regarded just as a 3D object. Indeed, similar change in slope due to the finite size effect is observed in percolation clusters [22]. We expect that, in much larger proteins, $\omega_{FS}$ shift towards the lower-frequency direction, and accordingly the region of $D_s \approx 1.3$ becomes wider. Thus we conclude that residue-residue interaction in proteins universally follows the power-law scaling $D_s$ with the spectral dimension $D_s \approx 1.3$.

We discuss the reason why some of the previous studies [14, 15] gave $D_s$ larger than 1.3. In these studies, $D_s$ was obtained not from DNM, i.e., the probability density function $\rho(\omega)$, but from its cumulative distribution function $\Omega(\omega) = \int_{0}^{\omega} d\omega' \rho(\omega')$. $D_s$ obtained from $\rho(\omega)$ is identical with that from $\Omega(\omega)$ if a single scaling holds over the whole range considered. In proteins, however, the scaling changes at around $\omega_{FS}$ due to the finite size
The gradient would yield $D_{s}$ of effect. This accordingly gives an illusionary larger value than

$$\omega_{FS}(s)$$

with a dimensionless positive constant $C$. Its cumulative distribution function is,

$$\Omega(\omega) = \int_{0}^{\omega} d\omega' \rho(\omega')$$

The gradient of log $\Omega$ to log $\omega$ gives a larger value than the correct spectral dimension $D_{s}$ at around $\hat{\omega} \gtrsim \omega_{FS}$. The gradient would yield $D_{s}$ in the region $\omega/\omega_{FS} \gg (1 - D_{s}/D)^{1/D_{s}}$. In proteins, $D = 3$ and $D_{s} = 1.3$, then $\omega/\omega_{FS} \gg 0.56$. This region, however, corresponds to the local motions, not to the global residue-residue interactions in which we have discovered the universality.

Conclusion: universality class of 3D critical percolation cluster. We have thus obtained the characteristic dimensions of FN inherent in the protein native structures, $(D, D_{c}, D_{f}, D_{s}) = (3, 1.9, 2.5, 1.3)$. Note that these dimensions are in surprisingly good coincidence with those in the 3D critical percolation cluster, $(D, D_{c}, D_{f}, D_{s}) = (3, 1.885, 2.53, 1.3)$ [21]. Hence we here propose that the protein native structures belong to the universality class of 3D critical percolation cluster. This is the main statement of this letter.

Then why proteins as residue-contact networks are critically percolated? Although it is difficult to give the complete answer in the present stage of this study, still we can provide a purposive explanation by pointing out two important aspects of proteins; stability and instability. On the one hand, proteins fold into their own (almost) unique native structures. Even when they are forced to unfold, they refold back into the native structures spontaneously (often with help from molecular chaperons). In this sense, proteins are stable. On the other hand, proteins flexibly change their structures. The structural change is sometimes accompanied with even (partial) unfolding. In this sense, proteins are unstable. The coexistence of these two conflicting aspects is essential for the functions of proteins, in particular to work as molecular machines or allosteric enzymes. Being in the critical state is sufficient for that. Furthermore, the criticality can be even necessary; proteins should evolve towards the critical state [26, 27]. This hypothesis should be verified through the study on molecular evolution, which is a challenging subject in the future.

This work was partially supported by Grants-in-Aids for Scientific Research in Priority Areas, the 21st Century COE Program (Physics of Self-Organization Systems), and “Academic Frontier” Project from MEXT.

---

[1] I. Bahar, A. R. Atilgan, and B. Erman, Folding & Design 2, 173 (1997).
[2] A. R. Atilgan et al. Biophys. J. 80, 505 (2001).
[3] H. Frauenfelder, S. Sligar, and P. Wolynes, Science 254, 1598 (1991).
[4] M. Vendruscolo et al. Phys. Rev. E 65, 061910 (2002).
[5] N. V. Dokholyan et al. Proc. Natl. Acad. Sci. USA 99, 8637 (2002).
[6] L. H. Greene and V. A. Itigam, J. Mol. Biol. 334, 781 (2003).
[7] A. R. Atilgan, P. Akan, and C. Baysal, Biophys. J. 86, 85 (2004).
[8] G. Bagler and S. Sinha, Physica A 346, 27 (2005).
[9] M. E. J. Newman, SIAM Rev. 45, 167 (2003).
[10] R. Solomonoff and A. Rapoport, Bull. Math. Biophys. 13, 107 (1951); P. Erdős and A. Rényi, Publ. Math. Debrecen 6, 290 (1959).
[11] D. J. Watts and S. H. Strogatz, Nature 393, 440 (1998).
[12] R. C. Herrick and H. J. Stapleton, J. Chem. Phys. 65, 4778 (1976); H. J. Stapleton et al. Phys. Rev. Lett. 45, 1456 (1980); J. P. Allen et al. Biophys. J. 38, 299 (1982); G. C. Wagner et al. J. Am. Chem. Soc. 107, 5589 (1985).
[13] J. S. Helman, A. Coniglio, and C. Tsallis, Phys. Rev. Lett. 53, 1195 (1984); R. Elber and M. Karplus, ibid. 56, 394 (1986).
[14] H. Wako, J. Phys. Soc. Jpn. 58, 1926 (1989).
[15] D. ben Avraham, Phys. Rev. B 47, 14559 (1993).
[16] X. Yu and D. M. Leitner, J. Chem. Phys. 119, 12673 (2003).
[17] M. B. Enright and D. M. Leitner, Phys. Rev. E 71, 011912 (2005); M. A. Moret et al. ibid. 71, 012901 (2005).
[18] G. Csányi and B. Szendrői, Phys. Rev. Lett. 70, 016122 (2004).
[19] http://www.rcsb.org/pdb/
[20] H. Morita and M. Takano, in preparation.
[21] D. Stauffer and A. Aharony, Introduction to percolation theory (Taylor and Francis, London, 1994), 2nd ed.
[22] T. Nakayama, K. Yakubo, and R. L. Orbach, Rev. Mod. Phys. 66, 381 (1994).
[23] F. Grassberger and I. Procaccia, Phys. Rev. Lett. 50, 346 (1983).
[24] M. Takano et al. Natural Computing 3, 377 (2004).
[25] D. A. Case et al. J. Comp. Chem. 26, 1668 (2005).
[26] S. A. Kauffman, Origins of order (Oxford university press, 1993).
[27] P. Bak, How nature works (Copernicus, New York, 1996).