PHYSICAL REVIEW E **79**, 020901(R) (2009)

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

Hidetoshi Morita[*] and Mitsunori Takano[†]
*Faculty of Science and Engineering, Waseda University, Tokyo 169-8555, Japan*

(Received 13 August 2008; published 24 February 2009)

Single protein molecules are regarded as contact networks of amino-acid residues. Relationships between the shortest path lengths and the numbers of residues within single molecules in the native structures are examined for various sized proteins. A universal scaling among proteins is obtained, which shows that the residue networks are fractal networks. This universal fractal network is characterized with three kinds of dimensions: the network topological dimension $D_c \approx 1.9$, the fractal dimension $D_f \approx 2.5$, and the spectral dimension $D_s \approx 1.3$. These values are in surprisingly good coincidence with those of the three-dimensional critical percolation cluster. Hence the residue contact networks in the protein native structures belong to the universality class of the three-dimensional percolation cluster. The criticality is relevant to the ambivalence in the protein native structures, the coexistence of stability and instability, both of which are necessary for protein functions.

DOI: 10.1103/PhysRevE.79.020901 PACS number(s): 87.15.B−, 05.45.Df, 64.60.−i, 89.75.Hc

Proteins are one-dimensional chains of amino-acid residues embedded in three-dimensional (3D) ($D=3$; 3D) Euclidean space. In a small scale we see covalent bonds of residues, and in a large scale we see just a 3D object. In the intermediate scale, we see the residues neighboring in the Euclidean space contacting with each other. Thus in this scale, referred to as the *network scale* in this paper, we can regard a single protein molecule as a contact network of residues [1,2]. This network viewpoint, in particular the universality among proteins, is complementary to the energy landscape picture [3] in understanding the general nature of proteins.

Some recent studies [4–8] have applied the latest network theory to the residue networks. Important quantities to characterize a network are the clustering coefficient $C$ and the shortest path length $L$ [9]. Those studies have demonstrated that the residue networks have larger $C$ than the random network [10], and smaller $L$ than the normal lattice. This indicates that the residue networks are small world networks (SWNs) [11].

On the other hand, the spacial profile of residues within single protein molecules has long been studied. Earlier spectroscopic studies have shown anomalous density of states [12]. These results, accompanied with theoretical studies [13], have suggested that the protein structures possess the property of fractal lattices. The fractality within single protein molecules has also been shown numerically through the density of normal modes [14–16] and the spatial mass distribution [17]. This implies that the residue networks are fractal networks (FNs).

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

In proteins, nevertheless, it is practically difficult to clearly distinguish between these two $N$ dependencies. 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 semilog (SWN) and log-log (FN) plots.

To overcome this practical 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 the $l$th path step. By overdrawing the $n_l$-$l$ plot for various sized proteins, we find a region of asymptotic universal scaling. We thereby conclude that the residue networks in the protein native structures are FNs, not SWNs. This is the first result of this paper.

We then obtain three characteristic dimensions of FNs; the network topological dimension $D_c$, the fractal dimension $D_f$, and the spectral dimension $D_s$. Their values are shown to be universal among single-chain proteins. Furthermore, these three values surprisingly coincide with those of the 3D critical percolation cluster. Namely, proteins belong to the universality class of the 3D critical percolation cluster. This is the second and the most highlighted result of this paper.

First of all, we define the network for a protein native structure. We use the spatial information of the native structure in the Protein Data Bank (PDB) [19]. We regard amino-acid residues as nodes; they are symbolized by $C_\alpha$ atoms, which is a standard way in coarse grained models [2], and is indeed employed in the past studies on networks [4,5,7,8]. A pair of nodes, $i$ and $j$, is considered to have an edge if their Euclidean distance $d_{ij}$ is less than a cutoff distance $d_c$. Then the network is represented by an adjacency matrix,

$$\mathbf{A} = (A_{ij}), \quad A_{ij} = \Theta(d_c - d_{ij}), \tag{1}$$

where $\Theta(\cdot)$ is the Heaviside step function. Here we adopt $d_c=7$ Å, which corresponds to the second coordination shell

[*] Present address: INLN, CNRS, 1361 route des Lucioles, 06560 Valbonne, France. Hidetoshi.Morita@inln.cnrs.fr

[†] mtkn@waseda.jp

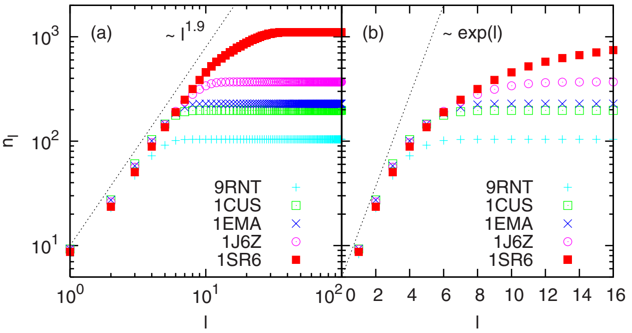

FIG. 1. (Color online) Averaged number of nodes $n_l$ that a walker on the network starting from a node can visit at least once until the $l$th step; (a) log-log and (b) semilog plots.

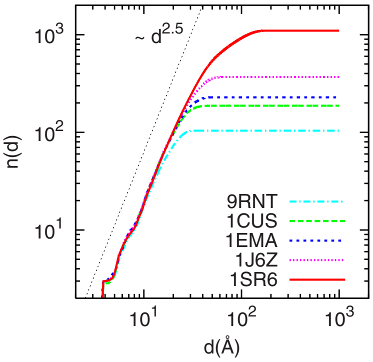

FIG. 2. (Color online) Averaged number of residues $n(d)$, the distance of which from a residue is less than $d$ for the same proteins as Fig. 1.

in the radial distribution function of $C_\alpha$; we have also confirmed that the result below is robust to the choice of $d_c$ from 6 to 10 Å [20].

Let $n_l^{(i)}$ be the number of nodes that a walker on the network starting from the node $i$ can visit at least once until the $l$th step. Since we are interested in the overall network properties of a protein molecule, we take the average, $n_l = \Sigma_i n_l^{(i)}/N$. As $l$ becomes larger, $n_l$ monotonically increases, and finally saturates at $N$. In the $D$-dimensional normal lattice, $n_l \sim l^D$. Similarly, in a FN, the power-law scaling,

$$n_l \sim l^{D_c}, \tag{2}$$

holds, where $D_c$ is referred to as the network topological dimension [21,22]. In a SWN, in contrast, the *small-world scaling* [18],

$$n_l \sim \exp(l/l_0), \tag{3}$$

holds, with a positive constant $l_0$. Note again that the relationships (2) and (3) are essentially different, leading to the dichotomy between FNs and SWNs [18].

Figure 1 shows $n_l$ vs $l$ in (a) a log-log and (b) a semilog plot. 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.). In Fig. 1(a), the scaling range tends to extend as the number of nodes $N$ increases. This suggests the existence of an asymptotic universal scaling in the network scale. In Fig. 1(b), on the contrary, we cannot see such an asymptotic behavior. Thus we infer that proteins as residue networks universally obey the power-law scaling (2) with $D_c \approx 1.9$. This provides evidence that the networks in protein native structures are FNs, not SWNs.

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

One plausible reason why the residue network is not a SWN but a FN is that the nodes are spatially restricted in the 3D Euclidean space, and therefore cannot have long-range edges. Long-range edges are important for SWN, as the Watts-Strogatz model [11] typically demonstrates. Indeed, some other real networks with spatial (geographical) restriction tend to be regular (including fractal) networks rather than SWNs [18].

Besides the above network topological dimension $D_c$, a FN is characterized in general by two other dimensions: the fractal dimension $D_f$ and the spectral dimension $D_s$ [21]. While these three dimensions and the Euclidean dimension $D$ are identical in the normal lattices, they can be different in FNs. We obtain the rest of the two dimensions in the following paragraph.

The fractal dimension is obtained from the spatial distribution of nodes. Here we again employ the method within single protein molecules, differently from the previous studies [17], in order to discuss the asymptotic scaling in the network scale. Let $n^{(i)}(d)$ be the number of nodes, the distance of which from the node $i$ is less than $d$; $n^{(i)}(d) = \Sigma_j \Theta(d - d_{ij})$. Since we are interested in the overall network property of a protein molecule, we take the average, $n(d) = \Sigma_i n^{(i)}(d)/N = \Sigma_i \Sigma_j \Theta(d - d_{ij})/N$. Note that this is nothing but the unnormalized correlation integral [23]. 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, in a FN, the power-law scaling,

$$n(d) \sim d^{D_f}, \tag{4}$$

holds, where $D_f$ is referred to as the fractal dimension.

Figure 2 shows $n(d)$ vs $d$ in log-log scale, for the same proteins as Fig. 1. Similarly to Fig. 1, the scaling range tends to extend as the number of nodes $N$ increases. This suggests the existence of an asymptotic universal scaling in the network scale. Thus we infer that proteins as residue networks universally obey the power-law scaling (4) with $D_f \approx 2.5$; this value is consistent with previous results [17].

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

$$\rho(\omega) \sim \omega^{D_s - 1}, \tag{5}$$

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

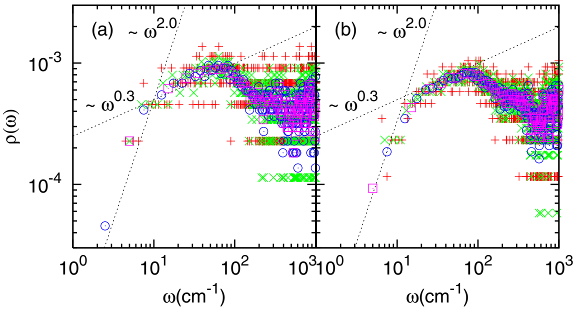

FIG. 3. (Color online) Density of normal modes $\rho(\omega)$ of (a) ribonuclease T1 (PDB ID=9RNT) and (b) cutinase (1CUS). Various bin sizes $\Delta\omega$ are taken [=1 (red plusses), 2 (green times symbols), 5 (blue circles), and 10 cm$^{-1}$ (magenta square)] so as to display the master curve more clearly.

DNM is, in general, obtained experimentally by spectroscopies and numerically by normal mode analysis (NMA). To be relevant to experiments, we conduct NMA in an all atom model, not in a coarse grained one. Then, by focusing on the frequency region corresponding to residue-residue interactions, 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 arbitrarily. It is true that coarse grained models well reproduce the overall fluctuation around 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 of frequency. 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 the AMBER force field (perm99) and implicit water (Generalized Born) models. Before NMA, energy minimization is executed with the Newton-Raphson and conjugate gradient methods, so that the norm of the force is less than the order of $10^{-12}$ kcal mol$^{-1}$ Å$^{-1}$.

Figure 3 shows typical DNMs among several proteins investigated; these are essentially similar to one of the previous studies [16]. There exist two shoulders at around 10 and 100 cm$^{-1}$, respectively, denoted by $\omega_{FS}$ and $\omega_{GL}$. The frequency higher than $\omega_{GL}$ corresponds to motions due to covalent-bond stretching and angle bending motions. The frequency lower than $\omega_{GL}$, in contrast, corresponds to motions due to residue-residue interactions that we are now interested in. In this region, i.e., the network region, DNMs universally obey the power-law scaling (5) with $D_s \approx 1.3$. At around $\omega_{FS}$, the dimension changes from 1.3 to 3.0. This is due to finite size effects; through a long wavelength probe, a protein is regarded just as a 3D object. Indeed, a similar change in slope due to finite size effects is observed in percolation clusters [22]. We expect that in much larger proteins $\omega_{FS}$ should shift to lower frequencies, and accordingly the region of $D_s \approx 1.3$ becomes wider. Thus we infer that proteins as residue networks universally follow the power-law scaling (5) with $D_s \approx 1.3$.

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 $\rho(\omega)$ but from its cumulative distribution $\Omega(\omega) = \int_0^\omega d\omega' \rho(\omega')$. $D_s$'s obtained from $\rho(\omega)$ and $\Omega(\omega)$ are identical if a single scaling holds over the whole range considered. In proteins, however, the scaling changes at around $\omega_{FS}$ due to finite size effects, which leads to an illusionary larger value of $D_s$. To illustrate this simply, we model the density function as a function that sharply changes the scaling at $\omega_{FS}$:

$$\rho(\omega) = \begin{cases} \dfrac{C}{\omega_{FS}} \left( \dfrac{\omega}{\omega_{FS}} \right)^{D-1} & (\omega \leqslant \omega_{FS}), \\ \dfrac{C}{\omega_{FS}} \left( \dfrac{\omega}{\omega_{FS}} \right)^{D_s-1} & (\omega > \omega_{FS}), \end{cases} \tag{6}$$

where $C$ is a dimensionless positive constant. Its cumulative distribution is

$$\Omega(\omega) = \begin{cases} \dfrac{C}{D} \left( \dfrac{\omega}{\omega_{FS}} \right)^{D} & (\omega \leqslant \omega_{FS}), \\ \dfrac{C}{D_s} \left[ \left( \dfrac{\omega}{\omega_{FS}} \right)^{D_s} - \left( 1 - \dfrac{D_s}{D} \right) \right] & (\omega > \omega_{FS}). \end{cases} \tag{7}$$

The gradient of log $\Omega$ to log $\omega$ gives a larger value than the correct spectral dimension $D_s$ at around $\omega \gtrsim \omega_{FS}$. 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 motions of covalent bonds, not to the motions of residue-residue interactions in which we have found the universality.

In conclusion, we have unveiled the existence of the universal FN inherent in the protein native structures in the network scale. We have obtained its characteristic dimensions $(D, D_c, D_f, D_s) = (3, 1.9, 2.5, 1.3)$. Note that these dimensions are in surprisingly good coincidence with those in the 3D critical percolation cluster, $(D, D_c, D_f, D_s) = (3, 1.885, 2.53, 1.3)$ [21,26]. Hence we here propose that the protein native structures belong to the universality class of the 3D critical percolation cluster. This is the main statement of this paper.

Then why are the residue networks critically percolated? Although it is difficult to give a complete answer in the present stage of this study, still we can provide a purposeful explanation by pointing out two important aspects of proteins: stability and instability. Proteins are stable, in that they keep their own almost unique native structures, which is necessary for robust molecular recognition in cells. Proteins are unstable, on the other hand, in that they change their structures flexibly, in particular to work as molecular machines or allosteric enzymes. The coexistence of these two conflicting aspects is essential for functions of proteins. Being in the critical state of the percolation transition is sufficient for fulfilling these two conflicting aspects. Indeed, recall that the percolation transition occurs as the density of nodes increases [21]. At low densities, the network is not percolated but parted, thereby the protein could not keep the native state. At high densities, the network is too much developed, thereby the protein would be too stiff to make a structural change. Furthermore, the criticality can be even necessary; proteins should evolve toward the critical state [27,28]. This hypothesis should be verified through molecular evolutionary studies, which is a challenging subject in the future.

The criticality in the percolation transition is likely to be relevant to the marginal stability of the native state in the folding-unfolding transition. Criticality means little curvature of free energy, which implies a shallow basin of and accordingly a low energy barrier from the native state. Such a free energy landscape can lead to large structural changes, including the folding-unfolding transition, without difficulty, which indicates the marginal stability of the native structure. In fact, it has been reported that free energy barriers to large scale structural change are quite low [29], which is consistent with the above picture of free energy as a consequence of the critical percolation that we have discovered in this paper.

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 the "Academic Frontier" Project from MEXT.

[1] I. Bahar, A. R. Atilgan, and B. Erman, Folding Des. **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, N. V. Dokholyan, E. Paci, and M. Karplus, Phys. Rev. E **65**, 061910 (2002).

[5] N. V. Dokholyan *et al.*, Proc. Natl. Acad. Sci. U.S.A. **99**, 8637 (2002).

[6] L. H. Greene and V. A. Higman, 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. Erdös and A. Rényi, Publ. Math. (Debrecen) **6**, 290 (1959).

[11] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).

[12] R. C. Herrick and H. J. Stapleton, J. Chem. Phys. **65**, 4778 (1976); H. J. Stapleton, J. P. Allen, C. P. Flynn, D. G. Stinson, and S. R. Kurtz, 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, J. G. V. Miranda, E. Noqueira, M. C. Santana, and G. F. Zebende, *ibid.* **71**, 012901 (2005).

[18] G. Csányi and B. Szendröi, Phys. Rev. E **70**, 016122 (2004).

[19] http://www.rcsb.org/pdb/

[20] H. Morita and M. Takano (unpublished).

[21] D. Stauffer and A. Aharony, *Introduction to Percolation Theory*, 2nd ed. (Taylor and Francis, London, 1994).

[22] T. Nakayama, K. Yakubo, and R. L. Orbach, Rev. Mod. Phys. **66**, 381 (1994).

[23] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).

[24] M. Takano *et al.*, Nat. Comput. **3**, 377 (2004).

[25] D. A. Case *et al.*, J. Comput. Chem. **26**, 1668 (2005).

[26] C. D. Lorenz and R. M. Ziff, Phys. Rev. E **57**, 230 (1998); P. N. Ballesteros *et al.*, J. Phys. A **32**, 1 (1999); Y. Deng and H. W. J. Blote, Phys. Rev. E **72**, 016126 (2005).

[27] S. A. Kauffman, *Origins of Order* (Oxford University Press, New York, 1993).

[28] P. Bak, *How Nature Works* (Copernicus, New York, 1996).

[29] O. Miyashita, J. N. Onuchic, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **100**, 12570 (2003).