# Small-world view of the amino acids that play a key role in protein folding

M. Vendruscolo,[1] N. V. Dokholyan,[2] E. Paci,[1,3] and M. Karplus[2,3]
[1]*Oxford Centre for Molecular Sciences, Central Chemistry Laboratory, South Parks Road, OX1 3QH Oxford, United Kingdom*
[2]*Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138*
[3]*Laboratoire de Chimie Biophysique, ISIS, Université Louis Pasteur, 4 rue Blaise Pascal, 67000 Strasbourg, France*
(Received 25 May 2001; revised manuscript received 21 March 2002; published 25 June 2002)

We use geometrical considerations to provide a different perspective on the fact that a few selected amino acids, the so-called "key residues," act as nucleation centers for protein folding. By constructing graphs corresponding to protein structures we show that they have the "small-world" feature of having a limited set of vertices with large connectivity. These vertices correspond to the key residues that play the role of "hubs" in the network of interactions that stabilize the structure of the transition state.

DOI: 10.1103/PhysRevE.65.061910 PACS number(s): 87.15.By, 64.60.Cn, 87.10.+e

Although proteins are complex systems, experiments [1] and theory [2,3] suggest that at least for some of them the folding mechanism is simpler than expected. One aspect of this is the finding that a small number of amino acids play an essential role in folding [4–8]. By applying the small-world network paradigm [9], we obtain a different perspective on this result and obtain a method for identifying the key amino acids.

Small-world networks [9–16] have recently been shown to be suitable for describing systems as diverse as chemical reaction networks [13], neural networks [9], food webs [14], social networks [9], scientific collaborations [12], disease spreading [15], and the World Wide Web [10,16]. In general, network topologies are random, if each vertex is connected randomly to other vertices, or they are regular, if each vertex is connected with a fixed number of vertices; two vertices are neighbors if they are connected by an edge. Watts and Strogatz [9] have shown that there exists a third possibility, corresponding to another regime of connectivity, which they called a *small-world* network. The networks that they describe are the result of the random replacement of a fraction $p$ of the edges of a $d$-dimensional regular lattice with new random edges. This results in connections between vertices that are distant on the lattice. The latter dramatically reduce the average path length $L$, where $L$ is equal to the number of vertices that must be traversed to reach any other vertex from a given one. Watts and Strogatz [9] characterized the small-world networks with two numbers, the average path length $L$ and the clustering coefficient $C$, which is the average fraction of pairs of neighbors that are also neighbors of each other. A vertex $k$ is connected to $N_k$ other vertices and the distribution $P(N_k)$ of the number of connections is either exponential, as in the original Watts and Strogatz model [9], or obeys a power law, as for example in the World Wide Web [16]. Regular networks have large $L$ and large $C$ whereas random networks have small $L$ and small $C$. Small-world networks have small $L$ and large $C$ [9]. In what follows we show that protein structures form small-world networks and use this result to identify key residues for the folding process [6]. The existence of key residues is in accord with the nucleation-condensation model of protein folding [4,5,8,17], in that they play an essential role in the folding nuclei [6,18]. The small-world character of networks in protein structures is shown to arise from the presence of a relatively small number of vertices with many connections [19,20].

To apply the small-world concept to an ensemble of protein structures we represent the latter as a weighted graph [21]. In order to do so, we first construct the adjacency matrix $\mathbf{A}$. The element $A_{ij}$ of $\mathbf{A}$ is given by the number of structures in which residues $i$ and $j$ are in contact divided by the total number of structures in the ensemble. For a particular structure, two residues are defined to be in contact if their $C_\alpha$ atoms are closer than a cutoff distance $R_c$ [6]. From the adjacency matrix $\mathbf{A}$ we construct the matrix of the distances $\mathbf{w}$ by defining its elements as $w_{ij}=1/A_{ij}$. For an individual structure, $w_{ij}=1$ if $i$ and $j$ are in contact and $\infty$ otherwise. In the general case, $1 \leqslant w_{ij} \leqslant \infty$. Each protein residue corresponds to a vertex of the graph and each element $w_{ij}$ corresponds to a weighted edge between two vertices. The *graph path length* $L$ is defined as

$$L=\frac{1}{N_p}\sum_{j>i}\lambda_{ij}, \tag{1}$$

where, in a graph of $N$ vertices, the sum runs over all the $N_p=N(N-1)/2$ pairs of vertices and $\lambda_{ij}$ is the *minimal path* between vertices $i$ and $j$. The minimal path $\lambda_{ij}$ is the minimum over all the paths between $i$ and $j$ of the sum of the weights of the edges traversed along each path. For graphs corresponding to individual structures we also defined the *clustering coefficient* $C$, as follows. If the vertex $k$ has $N_k$ neighbors, the maximal number of edges between the $N_k$ neighbors is $N_k(N_k-1)/2$. The clustering coefficient is

$$C=\frac{1}{N}\sum_{k}\frac{n_k}{N_k(N_k-1)/2}, \tag{2}$$

where $n_k$ is denoted by the actual number of edges that exist among the neighbors of $k$.

We determine the distribution of values of the path length $L$ and the clustering index $C$ for 978 representative protein structures from the Protein Data Bank (PDB) [22] whose sizes ranged from $N=50$ to $N=1021$. The result is shown in Fig. 1. The average value in the distribution for $L$ is $4.1 \pm 0.9$ and the average in the distribution of $C$ is $0.58 \pm 0.04$. If $N$ is the number of vertices and $K$ is the average number of

 

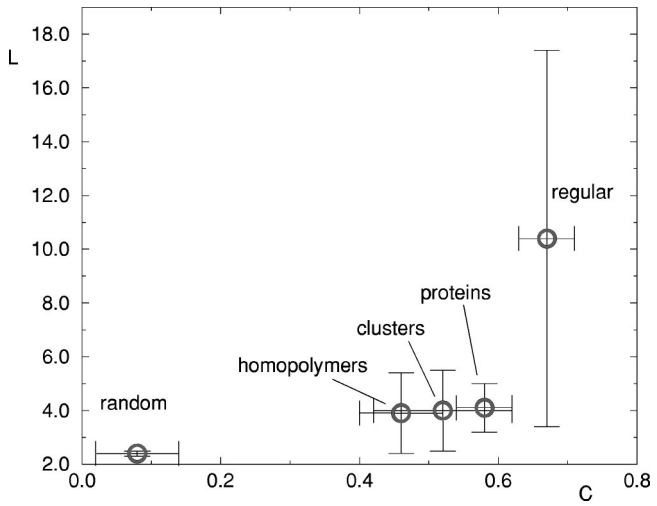


FIG. 1. Distribution of the values of the path length and clustering index for 978 representative proteins; for each one, a single structure from the PDB was used. Error bars represent the standard deviations of the distributions. For comparison, we also plot data points for random graphs, regular graphs, homopolymers, and atomic clusters. The conformations for homopolymers are obtained with the contact map dynamics of Ref. [27] and those of atomic clusters with Lennard-Jones interactions by a Monte Carlo method [28]. In the latter two cases, we considered sizes from $N=50$ to $N=1021$, a range comparable to that of single domain proteins.

neighbors in the graph, $L_{random} \sim \ln N / \ln K$ $(2.4 \pm 0.3)$ and $C_{random} \sim K/N$ $(0.08 \pm 0.06)$ for random graphs while for regular graphs (1 lattices, Ref. [23]), $L_{regular} = N(N+K-2)/[2K(N-1)]$ $(10.4 \pm 7.0)$ and $C_{regular} = 3(K-2)/[4(K-1)]$ $(0.67 \pm 0.04)$. The differences between the proteins and the random or regular lattices in Fig. 1 are statistically significant—according to the Kolmogorov-Smirnov test [24] the probability to observe the differences by chance is close to zero. These results show that the native protein structures are characterized by intermediate values of $L$ and $C$, and therefore, belong to the class of small-world graphs. Interestingly, individual collapsed structures of homopolymers and of clusters, for which the results are also shown in Fig. 1, have values of $L$ and $C$ that are similar to those of native protein structures. The differences in $L$ and $C$ between homopolymers, clusters, and proteins are probably not significant and may be due to the fact that somewhat different energy functions were used to model the various systems.

To determine the amino acid residues that make the most important contribution to generating the small-world network, we use the "betweenness" $B_k$, [25], defined as the number of pairs $(i,j)$ of vertices such that the shortest path between $i$ and $j$ passes through $k$, normalized by the total number of pairs. Figure 2 shows the $B_k$ values as a function of residue number $k$ for the native states and the transition state ensemble of six proteins. The former are based on x-ray or nuclear magnetic resonance structures and the latter are obtained by a Monte Carlo sampling procedure of Ref. [6]. In this method residue-specific protein engineering experimental results ($\phi$ values) [17] are interpreted in terms of the fraction of native contacts that each residue forms in the transition state and this information is used to bias the sampling of conformational space towards the region of the transition state. There is a correlation between $B_k$ and the square of the number of contacts of $k$; for $R_c = 8.5$ Å, the value used here, the correlation coefficient is about 0.8. Thus, $B_k$ measures the centrality of a residue and provides a correction to the use of the number of contacts for describing the structural relevance of a residue; i.e., the key residues are not necessarily the residues with the largest number of contacts [6].

For the transition states of all six proteins it is evident that there is a small number (between 2 and 4) of residues (or regions) that have large $B_k$ values and that outside these regions, the values are 0.1 or less. Analysis of the transition states of these proteins have shown that there are certain residues, called key residues, which are critical for forming the nucleus that encodes the overall native structure [6]. The key residues are indicated by small squares in Fig. 2. In all cases, they involve residues with large $B_k$. For five out of six proteins, they correspond to residues with the largest $B_k$. In the sixth (1aps), there are three key residues, all of which have large $B_k$. Two of them (11 and 94) are the largest $B_k$ in the given region and the third is in a region of large $B_k$ (residues 45–54) but is not the largest in that region. There is an additional region (residues 37–39) with $B_k$ greater than 0.15, which does not contain a key residue. It corresponds to strand $\beta_2$ (see Fig. 3), which is the most buried one in the native state. Experiments and the results of Ref. [6] indicate that this strand is partially formed in the transition state, although the interactions made by the residues in $\beta_2$ are not crucial for the nucleation process. It is likely that the chain can form the folding nucleus only if $\beta_2$ is near its native position. However, since it does not contain a key residue, the high $B_k$ value in the region 37–39 must be regarded as a false positive.

If we now examine the $B_k$ results for the native state (Fig. 2), it is clear that there is a significantly larger number of residues with high $B_k$ values. This is not surprising because only a portion of the native structure (i.e., the folding nucleus) is essentially formed in the transition state ensemble, so that the variations in the rest of the structure average out the high $B_k$ present in individual members of the ensemble (see also below). In the native state, fluctuations in the number of neighbors are small and such averaging does not occur. This leads to a larger number of high $B_k$ values. For example, in the protein AcP [6] (see Fig. 3), all of the five $\beta$ strands and the two $\alpha$ helices have a few residues that are central in the native state graph. However, the residues belonging to the $\alpha$ helices and those belonging to the $\beta_4$ strand lose their importance in the transition state graph (shown in Fig. 3), in accordance with the description of the transition state structure given in Ref. [6], where it was found that only strands $\beta_1$, $\beta_3$, and $\beta_5$ are relevant for the nucleation process.

Comparison of the native state and transition state results shows that it is possible to predict the key residues from a knowledge of the $B_k$ values of the latter, but not the former. The information is partially masked in the native state by the formation of the rest of the network that has both key and

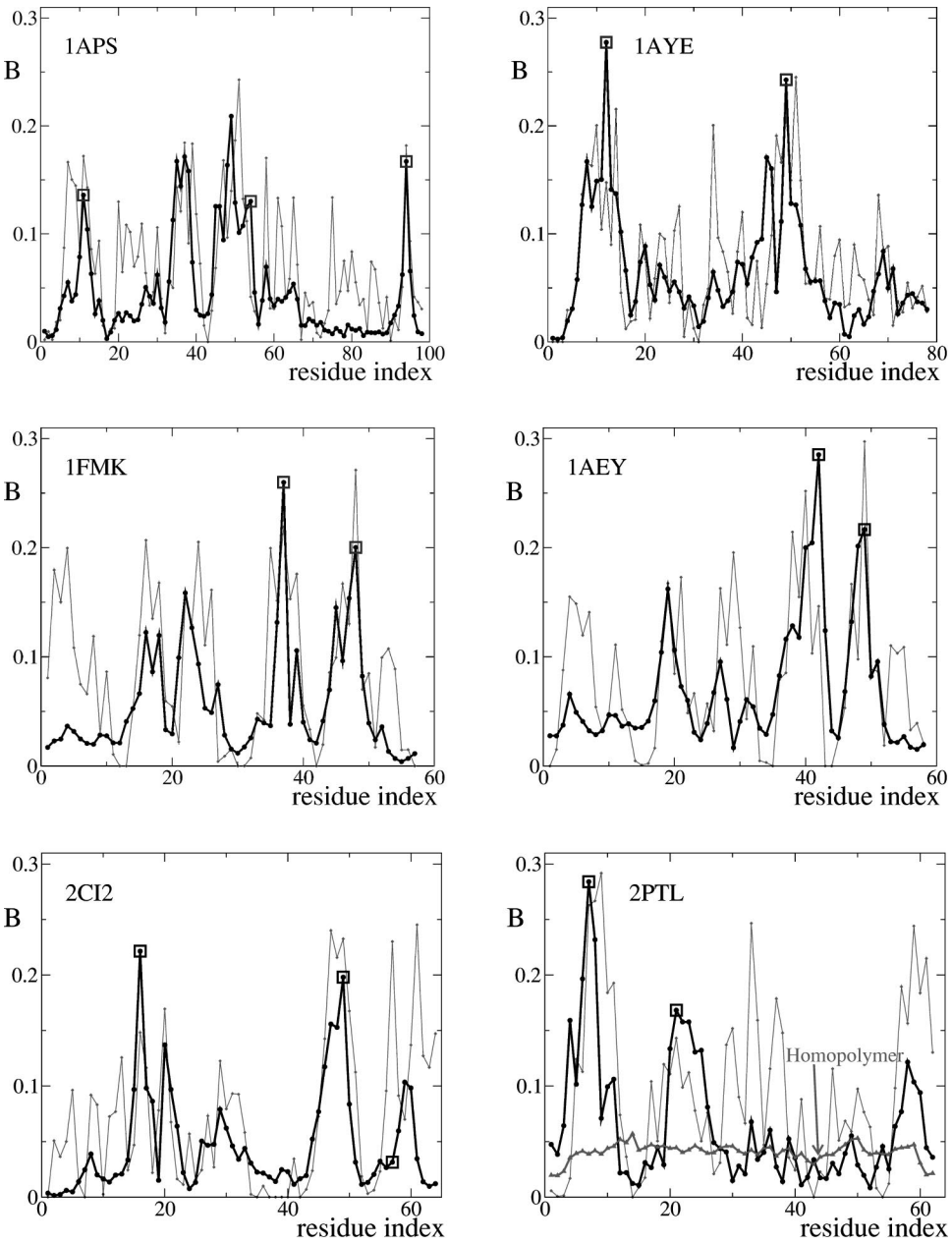


FIG. 2. "Betweenness" $B_k$ in the transition state for six proteins (thick lines). Vertices with large $B_k$ are the most connected ones. Key residues (obtained independently by the method presented in Ref. [6]) are indicated by squares. The $B_k$ values in the native state (thin lines) are shown for comparison. In the plot for protein 2ptl we show the $B$ profile for a homopolymer of the same length at the $\theta$ point (determined using the method of Ref. [29]).

non-key interactions. As a consequence, the small-world analysis of native states can be used to identify the regions in which key residues are expected to be found. However, the native state also identifies "false positives," namely, regions that are highly connected in the native state but not in the transition state. For example, in the case of AcP discussed above there are five candidate regions of which only three actually contain the key residues.

Individual compact structures of homopolymers and of atomic clusters have $B$ profiles similar to those of proteins of comparable size and their graphs have $L$ and $C$ values typical of small-world networks. This is due to the fact that we are dealing with systems of intrinsically finite size, so that in a collapsed polymer, a cluster or a globular protein, a relatively small number of residues are buried in the core and most are on the surface. Since the $B$ profiles are a measure of the average system connectivity, they are not very sensitive to the exact definition of contact. The similarity of the behavior of homopolymers and clusters suggests that chain connectivity, *per se*, plays only a minor role in this respect. The crucial difference between proteins and compact polymers is that the energy function of a protein selects one structure, that of the native state, with a non-negligible Boltzmann weight under native conditions. Instead for most homopolymers and clusters, a large number of compact conformations have similar probabilities. As a consequence, the $B$ profiles for homopolymers and clusters show no peaks when statistical averages are taken. This difference is found also when one compares the $\theta$ point for homopolymers and the transition states for proteins. As an example, we show in Fig. 2 the average $B$ profile for a homopolymer of the same length ($N=62$) as protein $L$ (2ptl). This difference is due to the fact protein folding takes place by a specific mechanism that involves few key residues selected by evolution. In this sense the


FIG. 3. Structure of the native state (left) and of the transition state (right) of the protein 1aps [6]. The contact network is shown in the transition state. The three key residues are indicated by spheres. Secondary structures ($\alpha$ helices and $\beta$ sheets) are indicated for the native state.

different order of the transition for protein folding and homopolymer collapse plays only a minor role. The difference between proteins and homopolymers is analogous to that between magic and nonmagic clusters. Magic clusters [26] are characterized by a single energy minimum whereas nonmagic clusters have a highly degenerate ground state. At low temperatures, therefore, there are geometrical key positions in a magic cluster. Due to symmetry under permutation, however, the identity of the atoms occupying them is not conserved. This situation is similar to that of homologous sequences with the same fold. The position in the structure is important, but the identity of the residues may change during evolution.

We have shown that structures of native proteins and of their transition states can be conveniently analyzed by using the small-world networks approach. Since this feature is also observed in collapsed homopolymers and in compact atomic clusters, it suggests that the small-world character arises primarily from the overall geometry (surface to volume ratio). What is special about proteins is that they have an essentially unique native structure and a structurally restricted ensemble representing the transition state. The betweenness in the transition state ensembles is highest for the key residues involved in formation of the nucleus for the folding reaction. It will be of interest to investigate whether the key residues identified in this way also play an energetic role in selecting the unique structure of the native state.

M.V. would like to thank EMBO for financial support. Work at Harvard (M.K.) was supported in part by a grant from the NIH. This work began while M.K. was Eastman Visiting Professor at Oxford University (1999–2000).

[1] D. Baker, Nature (London) **405**, 39 (2000).
[2] A.R. Dinner, A. Šali, L.J. Smith, C.M. Dobson, and M. Karplus, Trends Biochem. Sci. **25**, 331 (2000).
[3] A.R. Dinner and M. Karplus, Nat. Struct. Biol. **8**, 21 (2001).
[4] A.R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **92**, 10869 (1995).
[5] V.I. Abkevich, A.M. Gutin, and E.I. Shakhnovich, Biochemistry **33**, 10 026 (1994).
[6] M. Vendruscolo, E. Paci, C.M. Dobson, and M. Karplus, Nature (London) **409**, 641 (2001).
[7] F. Cecconi, C. Micheletti, P. Carloni, and A. Maritan, Proteins **43**, 365 (2001).
[8] N.V. Dokholyan, S.V. Buldyrev, H.E. Stanley, and E.I. Shakhnovich, J. Mol. Biol. **296**, 1183 (2000).
[9] D.J. Watts and S.H. Strogatz, Nature (London) **393**, 440 (1998).
[10] R. Albert, H. Jeong, and A.-L. Barabasi, Nature (London) **401**, 130 (1999).
[11] L.A.N. Amaral, A. Scala, M. Barthelemy, and H.E. Stanley, Proc. Natl. Acad. Sci. U.S.A. **97**, 11149 (2000).
[12] M.E.J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).
[13] U. Alon, M.G. Surette, N. Barkai, and S. Leibler, Nature (London) **397**, 168 (1999).
[14] K. McCann, A. Hastings, and G.R. Huxel, Nature (London) **395**, 794 (1998).
[15] M. Woolhouse and A. Donaldson, Nature (London) **410**, 515 (2001).
[16] A.-L. Barabasi and R. Albert, Science **286**, 509 (1999).
[17] A.R. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (Freeman, New York, 1999).
[18] E.I. Shakhnovich, V. Abkevich, and O. Ptitsyn, Nature (London) **379**, 96 (1996).
[19] M.E.J. Newman, J. Stat. Phys. **101**, 819 (2000).
[20] S.N. Dorogovtsev and J.F.F. Mendes, Europhys. Lett. **50**, 1 (2000).
[21] S.H. Yook, H. Jeong, and A.-L. Barabasi, Phys. Rev. Lett. **86**, 5835 (2001).
[22] H.M. Berman, J. Westbrook, Z. Feng, G. Gillirand, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[23] D.J. Watts, *Small Worlds. The Dynamics of Networks Between Order and Randomness* (Princeton University Press, Princeton, NJ, 1999).
[24] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes* (Cambridge University Press, Cambridge, U.K., 1989).
[25] L.C. Freeman, Sociometry **40**, 35 (1977).
[26] D.J. Wales and H.A. Scheraga, Science **285**, 1368 (1999).
[27] M. Vendruscolo and E. Domany, Fold Des **3**, 329 (1998).
[28] I. Andricioaei, J.E. Straub, and A.F. Voter, J. Chem. Phys. **114**, 6994 (2001).
[29] Y. Zhou, C.K. Hall, and M. Karplus, Phys. Rev. Lett. **77**, 2822 (1996).