Eur. Phys. J. B **49**, 127–134 (2006)
DOI: 10.1140/epjb/e2006-00026-0

THE EUROPEAN PHYSICAL JOURNAL B

The architectonic fold similarity network in protein fold space

Z.-B. Sun, X.-W. Zou[a], W. Guan, and Z.-Z. Jin

Department of Physics, Wuhan University, Wuhan 430072, P.R. China

Received 13 September 2005 / Received in final form 28 November 2005
Published online 31 January 2006 –

Abstract. By constructing the fold similarity network (FSN), we present an alternative approach to the characteristic and architecture of protein fold space. The degree distribution $P(k)$ of FSN differs far from that of the random network with the same number of nodes and connections. The investigation shows that FSN possesses small-world property and broad-scale feature. In order to access to the assumption of the dynamics behavior for FSN, we design a simple evolutionary dynamics model based on the duplication and variation fashions of protein folds. The simulation network generated by this model is a small-world one and reproduces the broad-scale degree distribution consistent with that of FSN. It seems that this model can be used to depict the divergent evolution and expanding progress of protein fold space.

PACS. 87.15.-v Biomolecules: structure and physical properties – 87.15.Aa Theory and modeling; computer simulation – 87.23.Kg Dynamics of evolution

1 Introduction

As the structural unit, protein domain has hydrophobic core and exists as a distinct region of protein 3D structure [1]. Therefore, many large proteins can be divided into spatially compact domains. It is well known that the domains can be clustered together into families based on their sequence identities of 30% or higher. Different families can be further grouped into folds according to their three-dimensional structure similarities [2]. If the domains have the same major secondary structures which are arranged with the same topology, they are defined as possessing a common fold. Consequently, many proteins with low sequence similarities share a similar three-dimensional structure or fold [3]. Since a number of sequences correspond to a specific fold, the number of folds is much smaller than that of sequences [4–6]. In addition, comparison of protein structures can reveal distant evolutionary relationship that would not be detected by sequence information alone. Therefore, the structure is a more robust characteristic for proteins comparing with sequence. With the greatly increasing in the number of structurally determined proteins, one of the principal goals of the structural genomics initiative is to enhance the understanding of protein fold space in the near future [7–12].

The first step in understanding complex 3D structures of proteins, deposited in Protein Data Bank (PDB) [13], is the classification of protein structures. Currently, there are several classification schemes that group the current set of known protein structures according to the structural similarities of their folds [14]. Each classification scheme is derived from different algorithms with slightly different goals. Among these classification schemes, the Structural Classification of Proteins (SCOP) [2] is a broadly used database for deriving the structural and evolutionary relationships of proteins. The entries in SCOP are protein domains. All entries can be classified hierarchically into families, superfamilies, folds, and classes. In addition, FSSP [15] is based upon the DALI algorithm [16,17], CATH [18] has its roots in the SSAP algorithm [19] and MMDB uses the VAST algorithm [20]. Anyway, through the comparison and classification of protein structures, one can understand the evolutionary and functional relationships among proteins [16,17,21–23].

In order to investigate the dynamic evolutionary progress of protein structure universe, Dokholyan et al. [24,25] constructed a structural similarity network named protein domain universe graph (PDUG). They considered only protein domains that had low pairwise sequence identities (<25%). On the basis of the rule to classify domains into families in SCOP, these domains are in principle equivalent to the SCOP families [2]. Therefore, PDUG could be regarded as a structural similarity network at the level of SCOP family. In PDUG, any pair of domains was connected on condition that the DALI Z-score [16,17], a measure of structural similarity, between them was above a given threshold value Z_{min}. With the aim of investigating protein structure universe at the level of family, they assigned $Z_{min} = 9$. As a result, these domains acting as family agents formed many disjoint clusters in PDUG, which were in principle equivalent to the SCOP folds [26]. They investigated the network characteristics of PDUG and built an evolutionary dynamics

[a] e-mail: xwzou@whu.edu.cn

model to reproduce the scale-free feature of PDUG. For the reason that each cluster in PDUG is disjointed, the network characteristics of PDUG are the sum of relationships within each cluster without regarding the relationships among the PDUG clusters, namely, the SCOP folds [26].

In protein universe, protein structures directly link to the biological functions thus are far more conserved than sequences in carrying phylogenetic signals [27,28]. It is believed that domains in one family have close evolutionary relationship while folds can pool distant relatives within different families. Therefore, folds are among the most conserved components in nature and can be used to study the distant evolutionary relationships [29]. Recently, Hou et al. [30] built a three-dimensional map of protein fold space based on the multidimensional scaling method. In their map, the structurally similar folds were represented by spatially adjacent points. It was found that all folds naturally separated into four distinct clusters, which corresponded to α, β, $\alpha+\beta$, and α/β classes in SCOP. The folds belonging to the same cluster had larger structural similarity, while the folds belonging to different clusters had smaller structural similarity. Going a further step, they designated a point as the evolution origin in the fold space. Moving away from the point, the size of the fold generally increased with increasing the length of its secondary structures and the folding complexity.

In order to outline more detailed relationships among protein folds, we construct a structural similarity network of folds (fold similarity network) and obtain some interesting characteristics of this network. With the attempt to reveal the principle underlying such characteristics and conjecture how the folds evolve during the biological evolution, we build a simple model based on the duplication and variation fashions of protein folds. The simulated result is quite suggestive that the currently observed folds may be originated from a divergent evolution of protein folds from one or a few precursor folds, such as the so-called 'evolution origin' in protein fold space.

2 Model and methods

2.1 Database

Despite the manual derivation, SCOP is a valuable structural classification database as the resource for detailed evolutionary information [2,14]. Hereby, we will build the data set of protein folds FD based on SCOP. The protein structures deposited in the Protein Data Bank may contain irregularities [31]. To overcome this shortage, the SPACI score is used as an approximate measure to report the quality of protein structures [32]. Based on this measure, the ASTRAL database selects a single representative domain for each SCOP fold according to the highest SPACI score [32]. We derive the representative domain for each fold from the ASTRAL database (`http://astral.berke ley.edu`) and form the fold data set FD. There are 179 α folds, 126 β folds, 121 α/β folds, and 234 $\alpha+\beta$ folds in SCOP 1.65, so FD consists of 660 domains. Each domain in FD represents a fold type.

By using the domains in FD, we expect to construct the fold similarity network (FSN) by the means analogous to PDUG [24]. It should be noted that, although PDUG and FSN both regard protein domains as nodes in graph or network, the notable distinction between them can be seen as follows. PDUG consists of all the domains that do not exhibit pairwise sequence similarity in excess of 25%. It means that each domain in PDUG is equivalent to a SCOP family, however, the domains in FSN represent the SCOP folds. It is well known that the families are highly unevenly distributed in folds [6]. There are 2051 families belonging to the 660 folds in FD. Among them, 384 (58%) folds contain only one family each, however, 66 (10%) superfolds contain 1057 (51.5%) families together. Herein, we use FD to represent the protein fold space.

2.2 Fold similarity network

In DALI program for structural alignment [16,17], Z-score is a quantitative measure of the structural similarity. Domany et al. [33] proposed that the matrix of pairwise Z-scores of domains could be viewed as a weighted graph and used to classify protein domains. In addition, Dokholyan et al. [24,25] obtained an unweighted protein domain universe graph by defining a threshold value of structural similarity, namely, $Z_{min} = 9$. In the present study we calculate the pairwise structural similarity of the 660 domains (folds) in FD by using the program DALI. Then we construct the structural similarity network of protein folds following others. Each node in the network expresses a fold, which is represented by the representative domain. Whether two nodes are connected or not depends on the structural similarity Z-score between the corresponding two folds. After assigning an appropriate threshold value of Z_{min}, any two folds in FD that have Z-score $Z \geq Z_{min}$ are connected by an identical edge (connection), otherwise, they are not. We create the fold similarity network (FSN) following this way. In FSN, each node denotes a fold in FD and each edge (connection) between two nodes represents that the structures of corresponding folds are similar.

As mentioned in former section, when $Z_{min} = 9$, PDUG consists of disjoint clusters. The network properties of PDUG are just the sum of relationships within each cluster. Therefore, the relationships among the PDUG clusters, in principle equivalent to the SCOP folds, are unexplored. As the domains in FD represent different folds in protein universe, we expect to uncover the relationships among the folds via the investigation of fold similarity network.

The study of network has recently become a blossoming area in science across many disciplines [34–41]. The properties of a network can be described by several parameters. In a network with N nodes [39,42,43], the degree k_v of node $v(v = 1, 2, \cdots, N)$ is the number of nodes that are connected to it. Hence, the average degree of a

network is $\langle k\rangle = \sum_v k_v/N$. The average degree $\langle k\rangle$, together with the total number of nodes N, can determine the size of the network. Besides these two parameters, the path length between two nodes is defined as the number of edges in the shortest path between them. The characteristic path length L of a network is the average path length over all pairs of nodes. In addition, the clustering coefficient C_v of node v is the connection fraction among k_v neighbors of node v, i.e., C_v is the ratio between the actual number of connections E_v and the possible number of connections $k_v(k_v-1)/2$. Thus we have $C_v = 2E_v/k_v(k_v-1)$. Therefore, the average clustering coefficient for the entire network is $C = \sum_v C_v/N$. In Section 3, we will discuss the network properties of FSN and compare them with those of a random network.

2.3 The evolutionary dynamics model

Recently, many models are developed to fit the broadly existing scale-free behaviors in complex networks [24,40,44]. The basic mechanism of these models is "preferential attachment" [38]. In biology, several duplication and divergence models have been proposed with preferential attachment to simulate the evolution of protein-protein interaction networks [45]. Analogously, we build an evolutionary dynamics model to reproduce the fold similarity network. In this model, each node represents a fold and the connection (edge) between two folds denotes that these two folds are structurally similar. In the beginning, we set up a single node to represent an initial fold. At each time step t, a new fold is generated by means of heredity or gene duplication, so the total number of folds is t after t steps. At the step t, we randomly select a fold v_m from the $t-1$ already existing folds as the parent fold and generate a new fold v_t as the offspring fold. The offspring v_t is structurally similar to the parent v_m, i.e., v_t and v_m are certainly connected. Whether the new fold v_t is connected to the neighbor folds of v_m or not is determined by the competition between heredity and variation. Thus, we introduce the variation threshold μ, which takes the value from 0 to 1. We can judge whether the variation of offspring fold v_t takes place or not by taking a stochastic number η ranging from 0 to 1. If $\eta \geq \mu$, the variation occurs and the offspring fold v_t is not connected to the neighbors of the parent fold v_m. If $\eta < \mu$, the offspring fold retains the heredity and has the opportunity to be connected to the neighbor folds of v_m. The connection probability is still μ. In order to determine whether the offspring fold v_t is connected to the neighbor v_i or not, we take another stochastic number $\eta^{'}$. If $\eta^{'} < \mu$, v_t and v_i are connected, otherwise, they are not. After N evolution steps, a simulation network with N nodes comes into being. It can be seen from the above discussion that, there is only one adjustable parameter in this model, the variation threshold μ, which determines the connections in the simulation network. After determining μ, the average degree $\langle k\rangle$ can be identified. Together with the number of nodes N, the size of the simulation network can be determined by a certain μ.

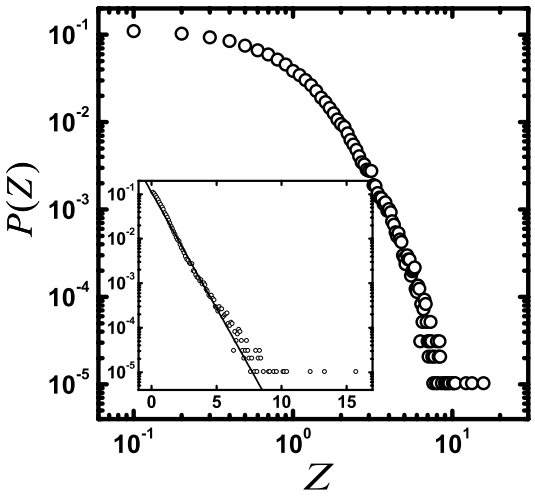

Fig. 1. The distribution of Z-scores in the fold data set (FD). Inset is the single-logarithmic plot of $P(Z)$. The full line is the plot of equation (1). It shows that the distribution of Z-scores follows an exponential relation.

To reproduce PDUG, Dokholyan et al. proposed a more complex model. In PDUG, each domain represented a sequence family and the families belonging to different folds constituted disjoint clusters. To be consistent with this, their model led to two situations: (1) the offspring and parent domains were structurally similar and connected (i.e., located in the same cluster and belonged to the same fold); (2) they were structurally dissimilar and not connected (i.e., located in different clusters and belonged to different folds). Therefore, their model depicted the evolutionary manner of protein families and reproduced the disjointed folds in PDUG. For the case of folds, FSN is a connected network, i.e., there is no disjointed clusters in FSN (see Sect. 3.1). Consequently, the offspring and parent folds are presumed to be always structurally similar and they are certainly connected in our model.

3 Results and discussion

3.1 Constructing FSN from FD

In order to construct the fold similarity network (FSN) from the fold data set (FD), we carry out all-against-all structural alignments for all folds in FD by DALI structure comparison algorithm [16,17]. Then we obtain 96 659 nonzero Z-scores from this procedure. The value of Z-score indicates the degree of structural similarity between two folds. In FD, the distribution of Z-scores $P(Z)$ as log–log plot is shown in Figure 1. The inset is a single-logarithmic plot. It shows that the distribution of Z-scores can be properly fitted by

$$P(Z) = 0.1\exp(-1.2Z). \quad (1)$$

Comparing Figure 1 with Figure 1b in the work of Dokholyan's group [25], we find that the distribution of

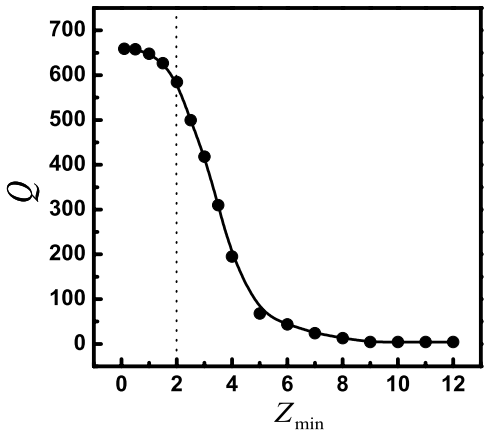

Fig. 2. The size of the largest connected network (Q) in FD against the threshold value Z_{min}. The dotted line represents $Z_{min} = 2$.

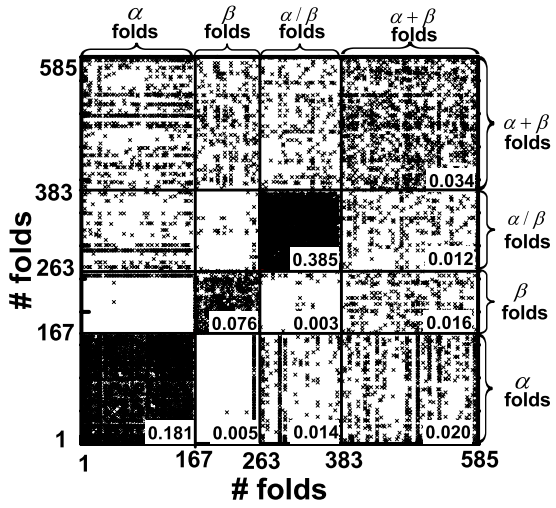

Fig. 3. Diagrammatic representation of the 585-dimensional symmetric adjacency matrix $\widehat{B}$ for the fold similarity network (FSN). The cross at site (i, j) represents that there is a connection between nodes v_i and v_j in FSN. In this matrix, the 585 folds are numbered in the sequence: $1 \sim 167$ (α), $168 \sim 263$ (β), $264 \sim 383$ (α/β), and $384 \sim 585$ ($\alpha+\beta$ fold class), respectively. In each class, the folds are numbered randomly. According to four fold classes, matrix $\widehat{B}$ can be partitioned into 10 different submatrices. The marked number in each submatrix is the density of connections in it.

Z-scores ranges to 10 for FD and to 40 for PDUG, respectively. This difference of Z-score distribution is due to the difference of domain selection. In the present work, we take one domain in each fold, and in PDUG the domains are randomly selected without considering their fold types. Therefore, the present distribution represents the behavior of folds and the corresponding distribution of PDUG indicates the characteristic of families. Next, we will determine the connections among folds according to the cutoff threshold Z_{min}.

It is generally accepted that, the aligned two structures have significant structural similarity only when $Z \geqslant 2$ [16,33,46,47]. For $Z_{min} = 2$, there are 585 (about 90%) folds in FD forming a connected similarity network. It indicates that nearly all folds in FD are structurally related, which implies the evolutionary relationship among the folds. Therefore, we naturally select the threshold value $Z_{min} = 2$. As a result, 585 folds in FD form the largest cluster by 7992 connections. Among the remaining 75 folds, 69 folds are orphan folds and 6 folds are connected in pairs. We can ascribe these scarce disconnected folds either to the result of serious variation in biological evolution or to the lack of the structural data disconnecting them from the largest cluster. After removing these rare variety, the 585 folds compose a connected fold similarity network (FSN). Among the 585 folds in FSN, there are 167 α folds, 96 β folds, 120 α/β folds, and 202 $\alpha + \beta$ folds, respectively.

To inspect the effect of different Z_{min} on the topological features of the similarity network, we calculate the number of folds Q belonging to the largest connected network in FD for various Z_{min}. The dependence of Q on Z_{min} is plotted in Figure 2. It can be seen that, when $Z_{min} > 2$, Q will decrease fast with Z_{min}. For instance, when $Z_{min} = 3$, only 418 (about 60%) folds in FD are included in the largest connected network and the rest are isolated. That is to say, about 1/3 folds have no structural and then evolutionary relationships to others, which would lead to an unreasonable similarity network. When $Z_{min} < 2$, large amount of trivial connections between the folds will be created in this work. For example, when $Z_{min} = 1$, the number of connections created above Z_{min} is 29 912. Because the created connections are identical, the trivial connections will blind the essential similarity among the folds. As a result, the degree distribution shows a maximum $P(k)$ around $k = 90$, which is similar to a random network. According to these reasons, we choose $Z_{min} = 2$ as the cutoff Z-score to define the connection between two folds in constructing the fold similarity network (FSN).

3.2 Representing FSN by the adjacency matrix

The fold similarity network (FSN) can be represented by a symmetric adjacency matrix $\widehat{B}$. The dimension of $\widehat{B}$ is 585, which corresponds to the number of folds in FSN. When two folds v_i and v_j are connected, the value of element b_{ij} in $\widehat{B}$ is 1, otherwise, it is 0. Figure 3 shows the adjacency matrix $\widehat{B}$ represented by symbols. When $b_{ij} = 1$, a cross appears at the site (i, j), otherwise, nothing occurs. In Figure 3, the 585 folds are numbered in the sequence of fold classes: $1 \sim 167$ (α), $168 \sim 263$ (β), $264 \sim 383$ (α/β), and $384 \sim 585$ ($\alpha + \beta$ fold class). In each fold class, the folds are arranged randomly. Therefore, the matrix $\widehat{B}$ can be partitioned into 10 different submatrices according to the fold classes. Among them, four submatrices along the diagonal of $\widehat{B}$ represent the intra-connectivities within α, β, α/β, and $\alpha + \beta$ folds, respectively. However, the other six submatrices represent the inter-connectivities between

two classes of folds. Therefore, we name these two kinds of submatrices as intra-submatrices and inter-submatrices.

It can be seen from Figure 3 that the distribution of symbols is uneven in $\widehat{B}$. The symbols are densely distributed in the four intra-submatrices, which indicates the high connectivity within α, β, α/β, and $\alpha+\beta$ folds, respectively. In each submatrix, the density of connections is defined as the ratio of the number of symbols (actual number of connections) to the number of sites (possible number of connections) in the submatrix. In Figure 3, the number marked in each submatrix is the corresponding density of connections. In the intra-submatrices, the density of connections is high. It means that the folds, which belong to the same fold class, have high probability to be structurally similar with each other. This result is well consistent with the fold map obtained by Hou et al. [30]. The density of connections is 0.385 for self-submatrix of α/β folds, however, it is 0.034 for that of $\alpha+\beta$ folds. Although these two classes of folds both consist of α and β secondary structures, they still have remarkable difference in the connectivities of intra-submatrices. The difference may arise from the distinct topological arrangement of α and β secondary structures in these two classes of folds. In α/β folds, α and β secondary structures are arranged alternative, which can induce the relative stable architecture and thus result in high probability to be structurally similar among α/β folds.

3.3 Structural characteristics of FSN

As described above, 585 folds in FD are connected and form the fold similarity network (FSN). In FSN, each node represent one fold and the connection between two nodes represents the structural similarity between the corresponding folds. The connections in FSN are unevenly distributed among the folds, which may imply the unique network properties of the FSN.

In general, the properties of a network can be characterized by several network parameters such as the degree, clustering coefficient, and characteristic path length. We calculate the degree k_v and clustering coefficient C_v for each node in FSN ($v = 1, 2, \cdots, 585$). The average degree $\langle k\rangle_{\rm FSN}$ is 27.3, average clustering coefficient $C_{\rm FSN}$ is 0.447, and the characteristic path length $L_{\rm FSN} = 2.8$. For comparison, we realize 100 random networks (RNs) with the same number of nodes and connections to FSN, i.e., $\langle k\rangle_{\rm RN} = \langle k\rangle_{\rm FSN}$. Differing from the connections in FSN, the connections in each realization of RN is randomly arranged. It is known that for random network consisting of N nodes, the average clustering coefficient $C_{\rm RN} = \langle k\rangle/N$ and the characteristic path length $L_{\rm RN} = \ln N/\ln\langle k\rangle$ [43]. Thus we can theoretically obtain that the average clustering coefficient $C_{\rm RN} = 0.046$ and characteristic path length $L_{\rm RN} = 1.9$ for random network (RN) with $N = 585$ and $\langle k\rangle = 27.3$. Meanwhile, the averaged values of $C_{\rm RN}$ and $L_{\rm RN}$ over the 100 realizations are 0.047 ± 0.001 and 2.23 ± 0.01, respectively. In graph theory, the two parameters L and C are customarily used to determine the small-world property of a network. Making a comparison between a network and a random network with the same number of nodes and connections, if the average clustering coefficient of the network $C \gg C_{\rm RN}$ and the characteristic path length $L \gtrsim L_{\rm RN}$, this network can be regarded as a small-world network [39]. Therefore, both the theoretic and calculated values shows that $C_{\rm FSN} \gg C_{\rm RN}$ and $L_{\rm FSN} \gtrsim L_{\rm RN}$, from which we can infer that FSN exhibits small-world property. The high clustering coefficient and small characteristic path length in FSN can be comprehended as the result of transition of the structural similarity in a cluster of folds, i.e., if a given fold has structural similarity to one of two structurally similar folds, it often has structural similarity to the other. Therefore, the neighbors of a fold tend to be connected to each other. It results in the high clustering coefficient for the folds in FSN. As to the characteristic path length, due to the continuous transition of the structural similarity, the path between two folds may become short. It leads to the small characteristic path length for FSN. As for the case of PDUG, because it consists of disjoint clusters, the two nodes belonging to different clusters can never be connected through any path. Thus, the path length between these two nodes is infinite. So does the characteristic path length of PDUG.

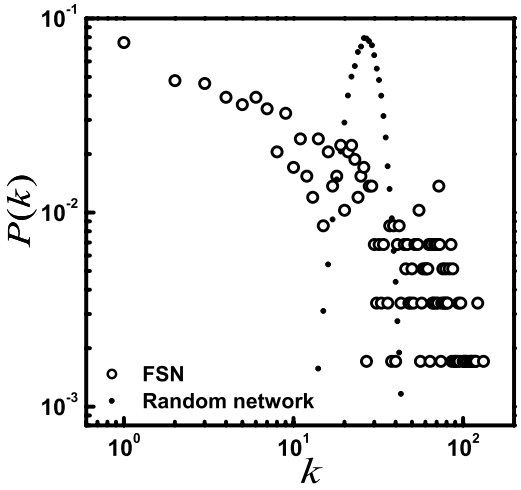

Fig. 4. The degree distribution of the fold similarity network (FSN) (in circles). Along with it is the degree distribution of the random network (in dots). For both networks, the node number $N = 585$ and the average degree $\langle k\rangle = 27.3$. The result of random network is averaged over 100 realizations.

To further investigate the properties of the fold similarity network, we calculate the probability $P(k)$ for the nodes that have degree k. The degree distribution of FSN is plotted in double-logarithm in Figure 4. For comparison, we also plot the degree distribution of the above random network (RN). It can be seen from Figure 4 that the degree distribution of RN has the form of a Poisson distribution with the peak around $\langle k\rangle = 27.3$. Differing dramatically from RN, the connected FSN has the highest probability for the folds possessing the lowest degree $k = 1$. It means that the dominant folds in FSN have only one connection to others. Figure 4 shows that in the region of small k, the

degree distribution has the form $P(k) = 0.3(k + 2.2)^{-1.1}$. This type of asymptotic power-law distribution is often found in wide fields of biology and physics, which indicates the "preferential attachment" as a general manner in the evolution of network [40]. In the region of large k, a cutoff takes place instead of power-law tail. The cutoff of the degree distribution at large k is often observed in real networks also [48]. The peculiar behavior of the degree distribution of FSN can be explained as follows. Firstly, we pay attention to the preferential attachment behavior at small k. If a fold has structural similarity to many other folds, it may contains commonly occurring super-secondary structural motifs (such as β-meanders, Greek keys, α-β plait motifs or α-hairpins) [49]. Therefore, such fold may have the opportunity to be structurally similar to more other folds. This is just the "preferentially attached" behavior: a new node prefers to be connected to the node which has already been connected to many nodes. Next we turn our attention to the cutoff behavior at large k. As a structural prototype, every fold has its unique architecture to tolerate the specific function. Therefore, a fold could not be structurally similar to too many folds. Even though a fold has the partiality for preferential attachment, the degree of such fold could not approach to an infinite value, which leads to the cutoff of degree distribution at large k region for FSN. Thus, FSN may be referred to as a broad-scale network [48].

3.4 The simulation of FSN with an evolutionary dynamics model

In the past subsection we have proved that the fold similarity network (FSN) has small-world property and can be classified as a broad-scale network. Now we want to ask: how does the fold space come into being, namely, how does the biological evolution imprint on the variation of folds? To address this question, we build an evolutionary dynamics model to simulate the formation of FSN based on duplication and variation fashions (see Sect. 2.3). In order to simulate the fold similarity network, the simulation network generated by this model should have the same size to the actual one, i.e., $N = 585$ and $\langle k\rangle = 27.3$. Therefore, we set the evolution step as 585 and intend to determine μ according to $\langle k\rangle = 27.3$. As a detail instruction, we will realize the relationship between μ and $\langle k\rangle$ at $N = 585$ by both the simulated and the theoretical means. In the simulation procedure, we obtain the simulated $\langle k\rangle$ with the error bar by averaging 100 realizations of this model for a given μ. The simulation relationship are plotted as dots in Figure 5.

Moreover, we derive the relationship between $\langle k\rangle$ and μ by a theoretical means. The procedure of derivation is as follows. Let K denote the sum of degrees of existing nodes in the network. At the time step t, an offspring node v_t is generated. If variation occurs, v_t is connected only to the parent node v_m but not to any neighbor of v_m. In this case, K is increased by 2 because only one connection is generated at this step. If variation does not occur, v_t is connected to both v_m and a part of neighbors of v_m.

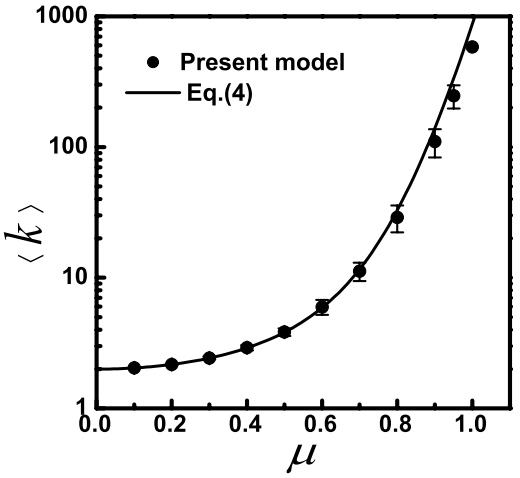

Fig. 5. The dependence of the average degree $\langle k\rangle$ on the variation threshold μ in present evolutionary dynamics model. The dots represent the simulation result and the curve is the theoretical result represented by equation (4). The number of nodes in the simulation network is $N = 585$.

Among k_m neighbors of v_m, which one is connected to v_t is stochastic, but the total number of connected neighbors is a certain value of μk_m. As a result, when variation does not occur, the increment of K is $2 + 2\mu k_m$. Because the probabilities of variation and unvariation are $1 - \mu$ and μ, respectively, the total increment of K is $2(1 - \mu) + (2 + 2\mu k_m)\mu$. Taking the average number of degree $\langle k\rangle$ $(=K/t)$ as the number of neighbors of the parent node k_m approximately, the increment of the sum of degrees of existing nodes from $t - 1$ to t time step can be written as

$$\frac{dK}{dt} = 2 + \frac{2\mu^2 K}{t}. \tag{2}$$

For any μ, when $t = 2$, $K = 2$. With this initial condition, equation (2) has a solution as

$$K = \frac{2t}{1 - 2\mu^2} + \left(1 - \frac{2}{1 - 2\mu^2}\right) \cdot 2^{1-2\mu^2} \cdot t^{2\mu^2}. \tag{3}$$

Substituting $\langle k\rangle$ $(=K/t)$ for K, and taking $t = N$, equation (3) becomes

$$\langle k\rangle = \frac{2}{1 - 2\mu^2} + \left(1 - \frac{2}{1 - 2\mu^2}\right) \cdot \left(\frac{N}{2}\right)^{2\mu^2 - 1}. \tag{4}$$

To be consistent with the size of FSN, we set $N = 585$. The theoretical results represented by equation (4) is plotted in Figure 5 too. Figure 5 shows that the theoretical result is consistent with the simulated one. It can be seen from Figure 5 that the average degree $\langle k\rangle$ depends monotonously on the variation threshold μ. As pointed in Section 3.3, the average degree $\langle k\rangle$ is 27.3 for FSN. In order to keep the same number of connections in present model with that in FSN, we should take $\langle k\rangle = 27.3$ in present model. Figure 5 shows when $\langle k\rangle = 27.3$, the corresponding variation threshold is about 0.8. Consequently, we can obtain the simulation network possessing the same size ($N = 585$

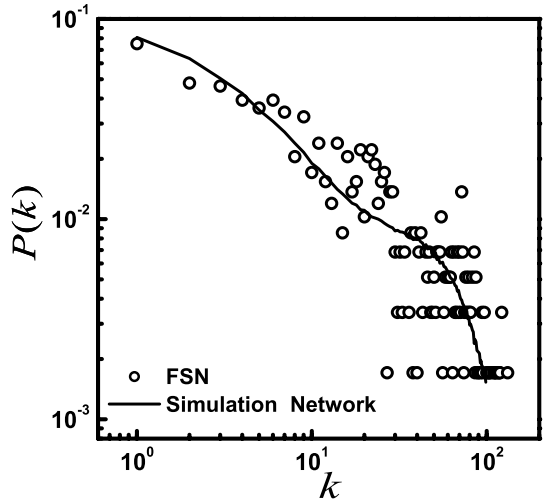

Fig. 6. The degree distribution of the simulation network (represented by real line) generated by the evolutionary dynamic model with $\mu = 0.8$. For comparison, the degree distribution of the FSN (represented by circles) is also plotted.

and $\langle k \rangle = 27.3$) to that of FSN by setting $\mu = 0.8$ in the present model.

To determine whether the simulation network resembles a small-world or a random one, we calculate its average clustering coefficients $C_{\rm EM}$ and characteristic path lengths $L_{\rm EM}$ and compare them with those of a random network. Over 100 realizations of the evolutionary model, the averaged values are $C_{\rm EM} = 0.561 \pm 0.025$ and $L_{\rm EM} = 3.69 \pm 0.40$. As to the random network with the same size, the corresponding values are $C_{\rm RN} = 0.047 \pm 0.001$ and $L_{\rm RN} = 2.23 \pm 0.01$. Thus, we obtain that $C_{\rm EM} \gg C_{\rm RN}$ and $L_{\rm EM} \gtrsim L_{\rm RN}$, which suggests that the simulation network possesses the small-world property. In addition, the broad-scale feature of the simulation network is determined by investigating its degree distribution. The simulated degree distribution averaged over 100 realizations of the present model is plotted in Figure 6 as real line. It shows that the simulation network also has the broad-scale feature. For comparison, we also re plot the degree distribution of the fold similarity network (FSN) in Figure 6. It can be seen that the degree distribution of the simulation network fits that of FSN reasonably. Therefore, besides the same number of nodes and the average degree as FSN, the simulation network possesses both the small-world property and broad-scale feature similar to that of FSN. Based on these results, we argue that the present model can reproduce the basic topological properties of FSN and thus be used to depict the divergent evolution and expanding progress of protein fold space.

It should be emphasis that the present model, being a schematic one, does not aim at a realistic description of protein evolution, which should be the result of the duplication and variation at the level of sequences or domains in a more complex way [3,24,25,50]. However, we still attempt to conjecture the scenario about the fold evolution by taking our eyesight on a higher level of protein fold. It seems that such model validates the viewpoint that all the modern proteins might have evolved from a few precursors.

4 Conclusions

Nature shows an extraordinary diversity of both sequence and structure while preserving biological function. This contradiction can be brought under a single roof if we notice that there is a limited number of protein folds in nature [49]. We construct the fold similarity network (FSN) and discover that FSN possesses of the small-world property and broad-scale feature. These characteristics of FSN may be derived from both the transitivity of structural similarity and the existence of super-secondary structural motifs in folds. FSN can be described by an evolutionary dynamics model based on duplication and variation fashions. It indicates that although the biological evolution are perfectly complex, the nature seems to economically invent new fold in a relatively simple manner. As a result, the arriving fold similarity network exhibits the elegant architecture.

We thank Dr. Liisa Holm for providing the DALI program. This work was supported by the National Natural Science Foundation of China No. 10374072 (X.W.Z.).

References

1. J. Janin, C. Chothia, Methods Enzymol. **115**, 420 (1985)
2. A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, J. Mol. Biol. **247**, 536 (1995)
3. N.V. Dokholyan, E.I. Shakhnovich, J. Mol. Biol. **312**, 289 (2001)
4. C. Chothia, Nature **357**, 543 (1992)
5. C. A. Orengo, D.T. Jones, J. Thornton, Nature **372**, 631 (1994)
6. X. Liu, K. Fan, W. Wang, Proteins **54**, 491 (2004)
7. C. Zhang, S. H. Kim, Curr. Opin. Chem. Biol. **7**, 28 (2003)
8. D. Baker, A. Sali, Science **294**, 93 (2001)
9. R. C. Stevens, I. A. Wilson, Science **293**, 519 (2001)
10. R. C. Stevens, S. Yokoyama, I. A. Wilson, Science **294**, 89 (2001)
11. R. Sanchez, U. Pieper, F. Melo, et al., Nat. Struct. Biol. **7**, 986 (2000)
12. B.E. Shakhnovich, N.V. Dokholyan, C. Delisi, E.I. Shakhnovich, J. Mol. Biol. **326**, 1 (2003)
13. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucleic Acids Res. **28** 235 (2000)
14. C. Hadley, D.T. Jones, Structure **7**, 1099 (1999)
15. L. Holm, C. Sander, Nucleic Acids Res. **26**, 316 (1998)
16. L. Holm, C. Sander, J. Mol. Biol. **233**, 123 (1993)
17. L. Holm, C. Sander, Science **273**, 595 (1996)
18. C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, J.M. Thornton, Structure **5**, 1093 (1997)
19. W.R. Taylor, T.P. Flores, C.A. Orengo, Protein Sci. **10**, 1858 (1994)
20. J.F. Gibrat, T. Madej, S. H. Bryant, Curr. Opin. Struct. Biol. **6**, 377 (1996)

21. W.R. Taylor, Nature **416**, 657 (2002)
22. I.N. Shindyalov, P.E. Bourne, Proteins **38**, 247 (2000)
23. J.F. Sadoc, R. Jullien, N. Rivier, Eur. Phys. J. B **33**, 355 (2003)
24. N.V. Dokholyan, B. Shakhnovich, E.I. Shakhnovich, Proc. Natl. Acad. Sci. USA **99**, 14132 (2002)
25. E.J. Deeds, N.V. Dokholyan, E. I. Shakhnovich, Biophys. J. **85**, 2962 (2003)
26. N. V. Dokholyan, Gene **347**, 199 (2005)
27. G. Caetano-Anollés, D. Caetano-Anollés, Genome Res. **13**, 1563 (2003)
28. A. Leo-Macias, P. Lopez-Romero, D. Lupyan, D. Zerbino, A.R. Ortiz, Biophys. Chem. **115**, 125 (2005)
29. N.V. Grishin, J. Struct. Biol. **134**, 167 (2001)
30. J.T. Hou, G.E. Sims, C. Zhang, S.H. Kim, Proc. Natl. Acad. Sci. USA **100**, 2386 (2003)
31. C.I. Branden, T.A. Jones, Nature **343**, 687 (1990)
32. S.E. Brenner, P. Koehl, M. Levitt, Nucleic Acids Res. **28**, 254 (2000)
33. G. Getz, M. Vendruscolo, D. Sachs, E. Domany, Proteins **46**, 405 (2002)
34. S.H. Strogatz, Nature **410**, 268 (2001)
35. A.R. Mashaghi, A. Ramezanpour, V. Karimipour, Eur. Phys. J. B **41**, 113 (2004)
36. M. Buchanan, *Small World: Uncovering Nature's Hidden Networks* (Weidenfeld Nicolson, London, 2002)
37. J. Weston, A. Elisseeff, D. Zhou, C.S. Lesie, W.S. Noble, Proc. Natl. Acad. Sci. USA **101**, 6559 (2004)
38. A.L. Barabasi, R. Albert, Science **286** 509 (1999)
39. D.J. Watts, S.H. Strogatz, Nature **393**, 440 (1998)
40. E.K. Koonin, Y.I. Wolf, G.P. Karev, Nature **420**, 218 (2002)
41. S. Wuchty, Mol. Biol. Evol. **18**(9), 1694 (2001)
42. S.N. Dorogovtsev, J.F.F. Mendes, Adv. Phys. **51**, 1079 (2002)
43. R. Albert, A.L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002)
44. J. Qian, N.M. Luscombe, M. Gerstein, J. Mol. Biol. **313**, 673 (2001)
45. J.S. Taylor, J. Raes, Annu. Rev. Genet. **38**, 615 (2004)
46. L. Holm, J. Park, Bioinformatics **16**, 566 (2000)
47. G. Getz, A. starovolsky, E. Domany, Bioinformatics **20**, 2150 (2004)
48. L.A.N. Amaral, A. Scala, M. Barthelemy, H.E. Stanley, Proc, Natl. Acad. Sci. USA **97**, 11149 (2000)
49. A. Harrison, F. Pearl, R. Mott, J. Thornton, C. Orengo, J. Mol. Biol. **323**, 909 (2002)
50. U. Bastolla, M. Vendruscolo, H.E. Roman, Eur. Phys. J. B **15**, 385 (2000)