BMC Structural Biology

Research article

Universal partitioning of the hierarchical fold network of 50-residue segments in proteins

Jun-ichi Ito[1], Yuki Sonobe[2], Kazuyoshi Ikeda[2,3,4], Kentaro Tomii[3] and Junichi Higo*[5]

Address: [1]Graduate School of Frontier Science, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8561, Japan, [2]School of Life Sciences, Tokyo University of Pharmacy and Life Sciences, 1432-1 Horinouchi, Hachioji, Tokyo, 192-0392, Japan, [3]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan, [4]PharmaDesign, Inc., 2-19-8 Hacchobori, Chuo-ku, Tokyo 104-0032, Japan and [5]The Center for Advanced Medical Engineering and Informatics, Osaka University, Open Laboratories for Advanced Bioscience and Biotechnology, 6-2-3, Furuedai, Suita, Osaka 565-0874, Japan

Email: Jun-ichi Ito - junichiito333@gmail.com; Yuki Sonobe - velvet_morning5@yahoo.co.jp; Kazuyoshi Ikeda - ikeda@pharmadesign.co.jp; Kentaro Tomii - k-tomii@aist.go.jp; Junichi Higo* - higo@protein.osaka-u.ac.jp

* Corresponding author

Published: 20 May 2009

BMC Structural Biology 2009, **9**:34 doi:10.1186/1472-6807-9-34

Received: 7 October 2008
Accepted: 20 May 2009

This article is available from: http://www.biomedcentral.com/1472-6807/9/34

Abstract

Background: Several studies have demonstrated that protein fold space is structured hierarchically and that power-law statistics are satisfied in relation between the numbers of protein families and protein folds (or superfamilies). We examined the internal structure and statistics in the fold space of 50 amino-acid residue segments taken from various protein folds. We used inter-residue contact patterns to measure the tertiary structural similarity among segments. Using this similarity measure, the segments were classified into a number (K_c) of clusters. We examined various K_c values for the clustering. The special resolution to differentiate the segment tertiary structures increases with increasing K_c. Furthermore, we constructed networks by linking structurally similar clusters.

Results: The network was partitioned persistently into four regions for $K_c \geq 1000$. This main partitioning is consistent with results of earlier studies, where similar partitioning was reported in classifying protein domain structures. Furthermore, the network was partitioned naturally into several dozens of sub-networks (i.e., communities). Therefore, intra-sub-network clusters were mutually connected with numerous links, although inter-sub-network ones were rarely done with few links. For $K_c \geq 1000$, the major sub-networks were about 40; the contents of the major sub-networks were conserved. This sub-partitioning is a novel finding, suggesting that the network is structured hierarchically: Segments construct a cluster, clusters form a sub-network, and sub-networks constitute a region. Additionally, the network was characterized by non-power-law statistics, which is also a novel finding.

Conclusion: Main findings are: (1) The universe of 50 residue segments found here was characterized by non-power-law statistics. Therefore, the universe differs from those ever reported for the protein domains. (2) The 50-residue segments were partitioned persistently and universally into some dozens (ca. 40) of major sub-networks, irrespective of the number of clusters. (3) These major sub-networks encompassed 90% of all segments. Consequently, the protein tertiary structure is constructed using the dozens of elements (sub-networks).

Background

Despite the vast number of amino-acid sequences, protein folds (or superfamilies) are quantitatively limited [1-4]. Consequently, protein fold classification is an important subject for elucidating the construction of protein tertiary structures. A key word to characterize protein folds is "hierarchy". Well-known databases – SCOP [5] and CATH [6] – have classified the tertiary structures of protein domains hierarchically. Similarly, a tree diagram was produced to classify the folds [7].

Mapping the tertiary structures of full-length protein domains to a conformational space, a structure distribution is generated: a so-called protein fold universe [8-11]. A key word to characterize the fold universe is "space partitioning". A two-dimensional (2D) representation of the fold universe was proposed in earlier reports [12,13], where the universe was partitioned into three fold (α, β, and α/β) regions. A three-dimensional (3D) fold universe was partitioned into four fold regions: all-α, all-β, α/β, and α+β [10]. Software that is accessible on a web site, PDBj http://eprots.protein.osaka-u.ac.jp/globe.cgi, serves the distribution on a global surface [14].

The structures of short protein segments have also been studied: Segments of a few (2–3) amino-acid residues long were projected in a two-dimensional (2D) space, where some typical combinations frequently appeared [15]. Fold universes of segments of 4–9 residues long [16] and 10–20 residues long [17-19] showed several clearly distinguishable structural clusters. A systematic survey for 10–50 residue segments has shown that the fold universe is classifiable into segment universes of three types: short (10–22 residues), medium (23–26 residues), and long (27–50 residues) [20]. In this work, the 3D shape of the universe varied abruptly at 23 and 27 residues long. A sequence-structure correlation found in short segments supports the tertiary structure prediction of full-length proteins [21-23].

These studies of protein segments and domains exemplify some structural clusters existing in the low-dimensional (2D or 3D) conformational space. The benefit of the low-dimensional expression is that one can readily imagine the shape of the universe. Increasing the segment length, however, the lowering of the space dimensionality hides the internal architecture of the structure distribution. Consequently, the internal architecture of the distribution for 50-residue segments (or longer segments) is unclear [20]. To compensate the full-dimensional information to the low-dimensional expression, a network is helpful in which two structures close to each other in the full-dimensional conformational space are connected.

Presume an ensemble of points (or nodes). Inter-node linkages form the networks. The network concept has been applied recently to biological systems [24-27]. Structurally similar segments can be linked for the segment fold universe. The structural similarity is computed for the overall structures of two segments (i.e., all coordinates of the segments). Therefore, the similarity is a quantity defined in full-dimensional space. Consequently, a 2D or 3D universe consisting of linked nodes involves full-dimensional information. To assign inter-node linkage in the ensemble, a score is important to quantify the structural similarity between two tertiary structures. Inter-residue contact (native contact) patterns have been used as reaction coordinates in protein folding studies [28-30]. When two structures have similar native contact patterns, they exhibit similar inter-residue packing. Results of several studies indicate that the native contacts are useful indicators to assess the protein folding process [31-43] and folding time scale [41-43].

Herein, we constructed a fold network of 50-residue segments taken from four major structural classes of protein domains. We used the inter-residue contact pattern for the similarity score. The resultant networks showed the main partitioning, as expected. Furthermore, as a new finding, the network of the segment structures was partitioned into dozens of universal communities (sub-networks). From these observations, we propose a novel protein structure hierarchy with community sites at a hierarchy level. The novelty of the currently identified hierarchy was ensured by non-power-law statistics in the hierarchy, which differs from power-law statistics characterizing other hierarchies ever found for protein tertiary structures.

Results

As described in *Methods*, 50-residue segments were taken from representative proteins and classified into K_c clusters, each of which consists of structurally similar segments. We calculated the native contact patterns that are common in each cluster, and constructed networks by connecting the clusters according to their contact pattern similarity. In *Results*, we first examine the general aspects of the obtained clusters. Second, we check the conformational distribution using a 3D map. Finally, we analyze the characterization of 50-residue segment universe using a network analysis.

As described in this paper, indices i and j are used for specifying residue positions in a 50-residue segment, s and t for segment ordinal numbers, u and v for cluster ordinal numbers, and w for a community ordinal number.

General aspects for clusters

Figure 1A portrays the dependence of the average cluster size $<S>$ (Eq. 3) on the number K_c of clusters. Actually, K_c determines the spatial resolution to view the universe of the 50-residue segments: With decreasing K_c, $<S>$ increases because structurally different segments are fused

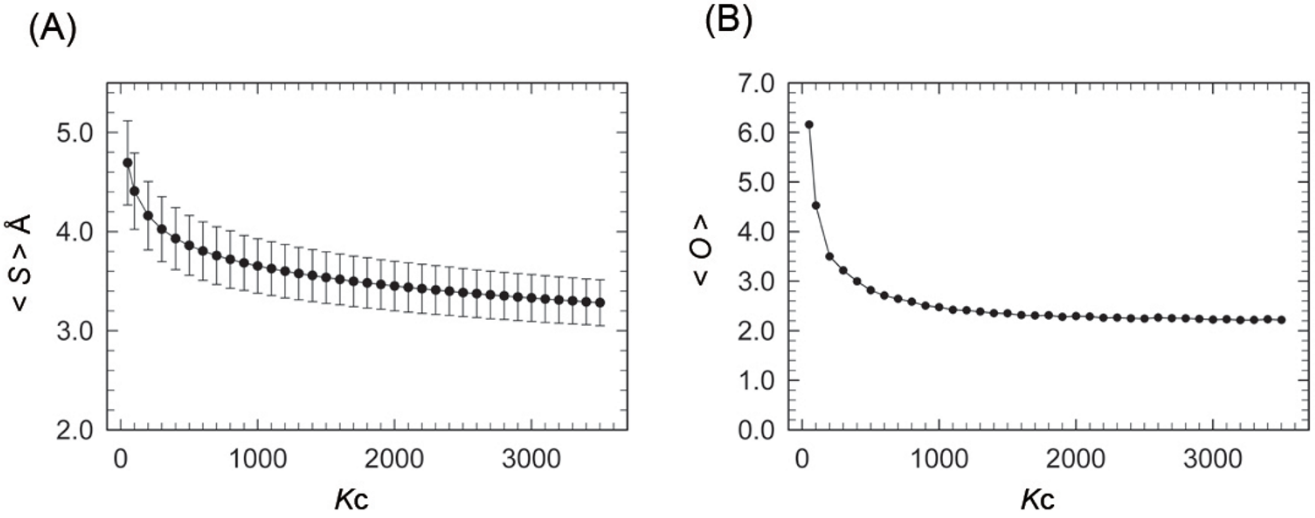


Figure 1
<*S* > and <*O* > as a function of *K*$_c$. (A) <*S* > is the average cluster size (Eq. 3). The error bar shows the standard deviation over clusters. (B) <*O* > is the average number of segments supplied by a protein to a cluster (see the text for a detailed definition of <*O* >).

into a cluster. The change of <*S* > was rapid for small K_c and slow for larger K_c.

The segments were generated by sliding a 50-residue window one residue by one residue along the domain sequences (see *Methods*). Consequently, two segments taken from the same protein domain with mutual adjacency in the sequence might have similar structures and might therefore be involved in a cluster. We did the following analysis to verify this possibility quantitatively: Presume that a cluster u involves n_m segments originated in a protein m. Subsequently, we introduced a quantity: $O_u = \sum_m n_m / N_\mathrm{p}$, where the summation is taken over proteins that supply segment(s) to the cluster u, and N_p is the number of those proteins. Figure 1B presents a plot of the average of O_u as a function of K_c: $< O >= \sum_u^{K_\mathrm{c}} O_u / K_\mathrm{c}$. For K_c = 1000, <*O* > converged to 2.2. Consequently, a protein supplies only two or three segments to a cluster on average: i.e., a cluster does not contain excessive segments derived from a single protein for $K_\mathrm{c} \geq 1000$.

Figure 2 depicts the number (n_u) of segments involved in a cluster as a function of the cluster ordinal number for K_c = 1000. The decay of n_u is non-exponential. It is particularly interesting that even cluster #950 involves more than 100 segments, which means that the cluster comprises more than 40 (= 100/2.5) different proteins (<*O* > ≈ 2.5 for K_c = 1000). In the last 50 clusters, n_u decreased quickly. These clusters consist of randomly structured segments. Although segments were taken from all-α, all-β, α/β, and α+β SCOP classes, the structures can be random.

Figure 3 depicts $<f>_{Kc}$ (Eq. 9) depending on K_c. The value of $<f>_{Kc}$ was 0.60–0.65 for $K_\mathrm{c} \geq 1000$. The similarity threshold f_0 for assigning the inter-cluster linkage (Eq. 7) was 0.7. Figure 3 presents that the inter-residue similarity is compatible with the intra-cluster similarity.

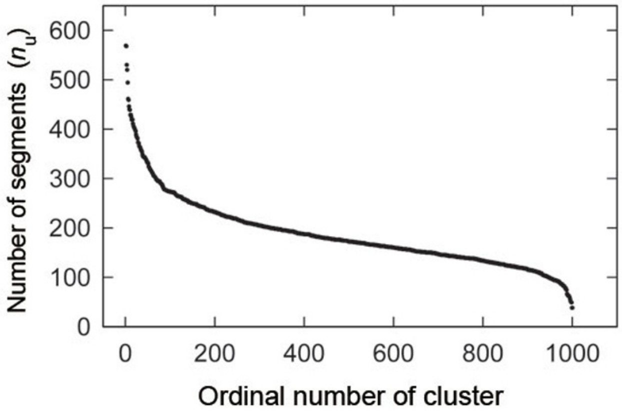


Figure 2
Number n_u of segments in a cluster as a function of the ordinal number of the cluster.

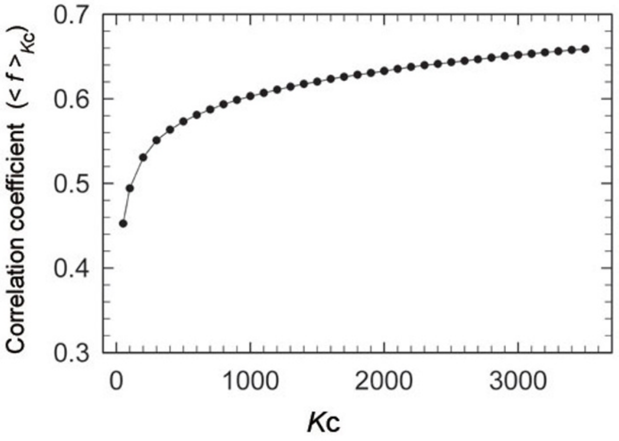


Figure 3
Averaged correlation coefficient $<f>_{Kc}$ (Eq. 9) for intra-cluster segments as a function of K_c.

Fold universe and network of clusters

The inter-cluster (inter-node) links were assigned to the K_c clusters according to the adjacency matrix a_{uv}. Directly connected clusters have mutually similar inter-residue contact patterns. Internal architectures of the networks were investigated by dividing the networks into communities (sub-networks) using Newman's method [44]. In parallel, we projected the networks into a 3D space to obtain positions in the conformational space (see Additional file 1 for details). Although the clusters were embedded in the 3D space, the inter-cluster links were given to clusters that are mutually close in the full-dimensional space.

Each community was characterized by five biophysical structural features: the α, β, αβ secondary-structure elements, the radius of gyration, and the number of inter-residue contacts, denoted respectively as n_α, n_β, $n_{\alpha\beta}$, R_g, and N_{contact}. Then, the communities were classified into four types (α, β, αβ, and randomly structured communities) depending on the five structural features (see *Methods* for details).

Figure 4 portrays the 3D cluster distributions at $K_c = 1000$, 2000, and 3000, where a single color was assigned to a community depending on secondary-structure elements n_α, n_β, and $n_{\alpha\beta}$ (see Additional file 1 for details). This figure clearly illustrates that the 3D cluster network is partitioned into four fold-regions (mainly α, mainly β, αβ, and randomly structured regions) independent of K_c, which respectively consist of α, β, αβ, and randomly structured communities. We termed this partitioning as "main partitioning". Figure 5 shows that the overall shape of the network adopted a three-leaf clover shape (mainly α, mainly β, and αβ regions surrounding the randomly structured region). We checked quantitatively whether the 3D distribution reflected the original full-dimensional distribution by calculating F-measure $\bar{F}_{\max}$ (see Additional file 1 for the definition of $\bar{F}_{\max}$). The value of $\bar{F}_{\max}$ was, respectively, 0.804 for $K_c = 1000$, 0.673 for $K_c = 2000$, and 0.593 for $K_c = 3000$. The large value of $\bar{F}_{\max}$ for $K_c = 1000$ indicates that the 3D cluster distribution fairly reflects the full-dimensional distribution. The $\bar{F}_{\max}$ value decreased con-

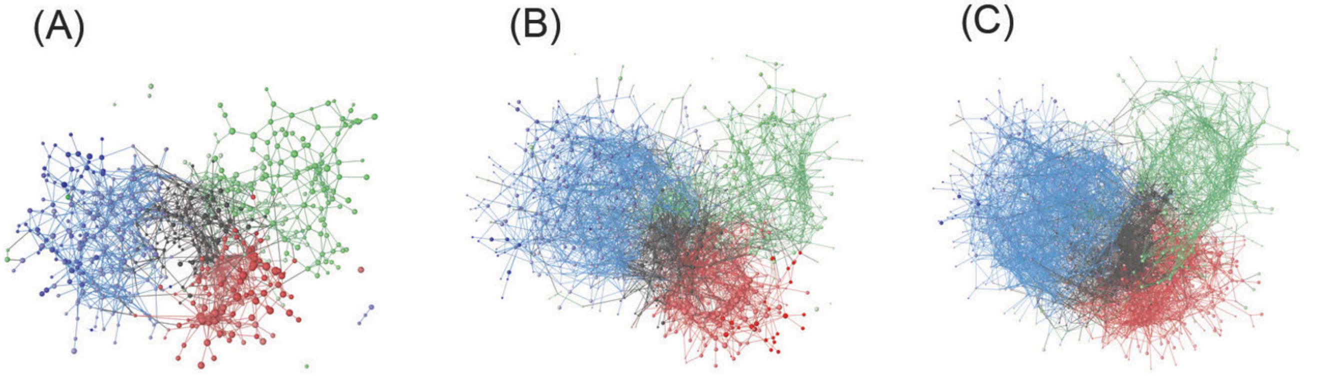


Figure 4
Networked 3D distribution of clusters for K_c = 1000 (A), 2000 (B), and 3000 (C). In this figure, a sphere represents a cluster. The larger the sphere, the more segments the cluster involves. The coloring method for clusters and inter-cluster links is explained briefly below (see Additional file 1 for details): The α, β, and αβ communities are, respectively, red, blue, and green. The larger the secondary-structure contents in a community, the greater the color strength. All randomly structured communities are shown in black. Colors assigned to cluster-cluster links are as follows: red for links within α communities, blue for those within β communities, green for those within αβ communities, and black for other links.

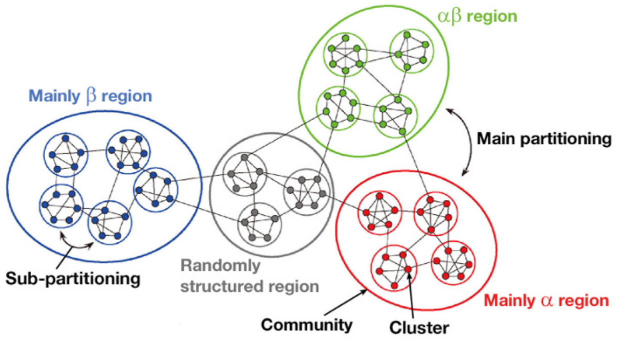


Figure 5
Main and sub-partitioning of the cluster network.

comitantly with increasing K_c. However, the three-leaf clover shape of the distribution was conserved at various K_c, which strongly suggests that the main partitioning exists in the 50-residue segments universe.

Figure 6 displays segment tertiary structures picked from clusters. This figure portrays that the structure classification by the five structural features correlates well with the visual secondary-structure constitution. Most segments originating in the all-α SCOP fold class were assigned to the α communities (see a-1 and a-2 in Figure 6). Those that originated in the all-β SCOP fold class were assigned to the β communities (see b-1 – b-3). The majority of segments taken from the α/β SCOP fold class were assigned to the αβ communities (see c-1 – c-4), although some were involved in other fold regions. In contrast, segments from the α+β SCOP fold class scattered to all the fold regions because the α+β proteins are a mixture of helices, strands, and randomly structured fragments, where the α and β secondary-structure elements are not necessarily neighbors to each other in the sequence. Consequently, the 50-residue segments from the α+β proteins can involve various structural features. The randomly structured region contained clusters with a few secondary-structure elements (see r-1 – r-4 in Figure 6). However, its polypeptide packing was loose, as portrayed in Figure 7, where the randomly structured clusters had large R_g.

Non-power-law statistics

The protein-domain universe is known to be an extremely biased distribution [8,45]. Many studies have suggested a power-law statistic to represent the relation between the number of families and the number of folds [9,46,47]. For instance, Shakhnovich and co-workers created a protein-domain universe graph (PDUG) with adoption of a DALI

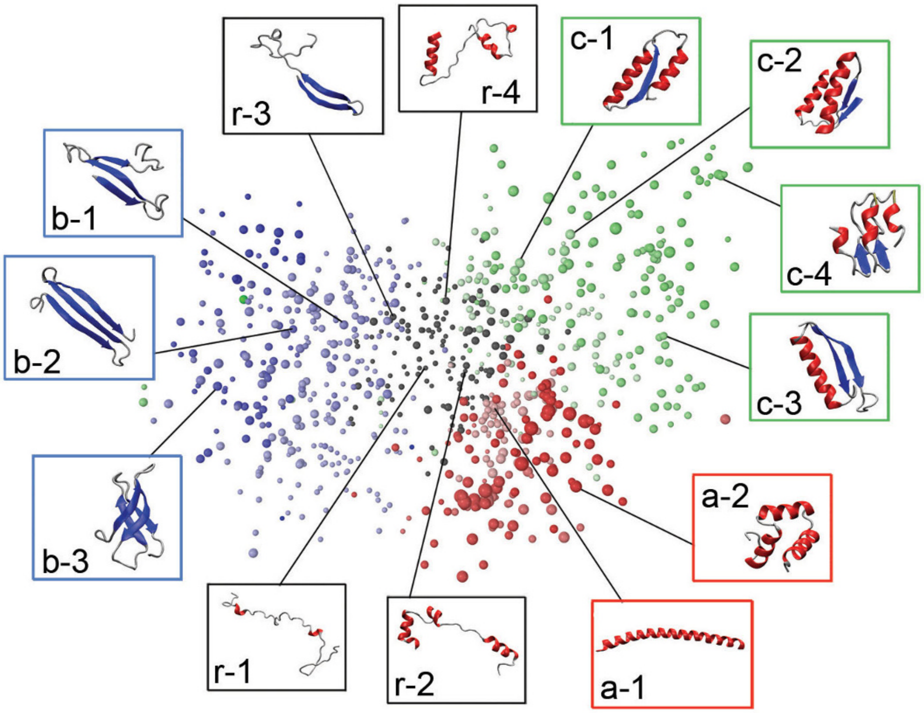


Figure 6
Tertiary structures picked from 3D distribution for K_c = 1000 Colors. of clusters are the same as those depicted in Figure 4. Inter-cluster links are not shown. This figure is presented with the same orientation as that of Figure 4.

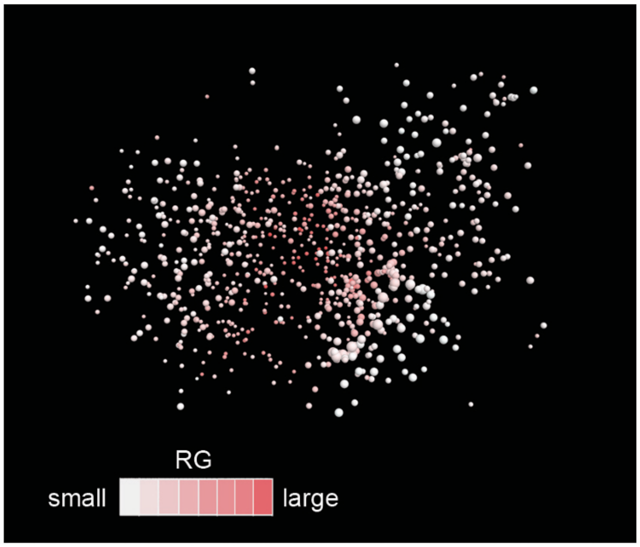


Figure 7
Radius of gyration R_g of clusters. With increasing R_g, the cluster color is redder. This figure is presented with the same orientation as that of Figure 4.

Z-score for the similarity score, and showed that the domain universe followed a power-law distribution [9]. Consequently, it is interesting to check if the currently produced network of the 50-residue segments follows the power law distribution.

First, we calculated the number (n_{seg}) of segments involved in each cluster. Figures 8A, B, and 8C portray the relation between n_{seg} and the number of clusters that respectively involve n_{seg} segments at K_c = 1000, 2000, and 3000. The distributions were symmetric (the value of skewness was 0.138 for K_c = 1000, 0.006 for K_c = 2000, and -0.066 for K_c = 3000) on the X-axis, log(n_{seg}), and far from the power-law statistics. Therefore, the currently obtained universe differs from those that have ever been reported. Additionally, we calculated the number (n'_{seg}) of segments involved in each community, and showed the relation between n'_{seg} and the number of communities involved n'_{seg} fragments for K_c = 1000, 2000, and 3000. We again obtained non-power-law statistics in the relation (data not shown).

Next, we calculated a connectivity distribution, $P(k)$, of the networks to investigate details of the cluster network [48]. The $P(k)$ is defined as a distribution function of clusters that have k links to other clusters. Figures 9A, B, and 9C respectively present $P(k)$ at K_c = 1000, 2000, and 3000. Subsequently, $P(k)$ decays exponentially with increasing k. Therefore, these distributions are exponential ones (or possibly truncated power-law distributions). Consequently, non-power-law networks (i.e., non-scale-free networks) are again observed for the current networks.

Robustness of communities

We conducted modularity analysis to study cluster networks from another perspective. First, the networks were divided into communities (see *Methods*). A modularity Q_{mod} is an index to assess how well the network is divided into communities [49]: $0 \leq Q_{\mathrm{mod}} \leq 1$. A network with a large Q_{mod} is characterized by numerous intra-community links and a few inter-community links. Figure 10A portrays the K_c dependence of Q_{mod}, which has the maximum at K_c = 200, indicating that the communities were highly isolated at K_c = 200. For K_c > 200, the communities were connected gradually by links, thereby decreasing Q_{mod}. For $K_c \geq 1000$, Q_{mod} converged to a value (0.63), which indicates that the 50-residue segment network is characterized by high modularity.

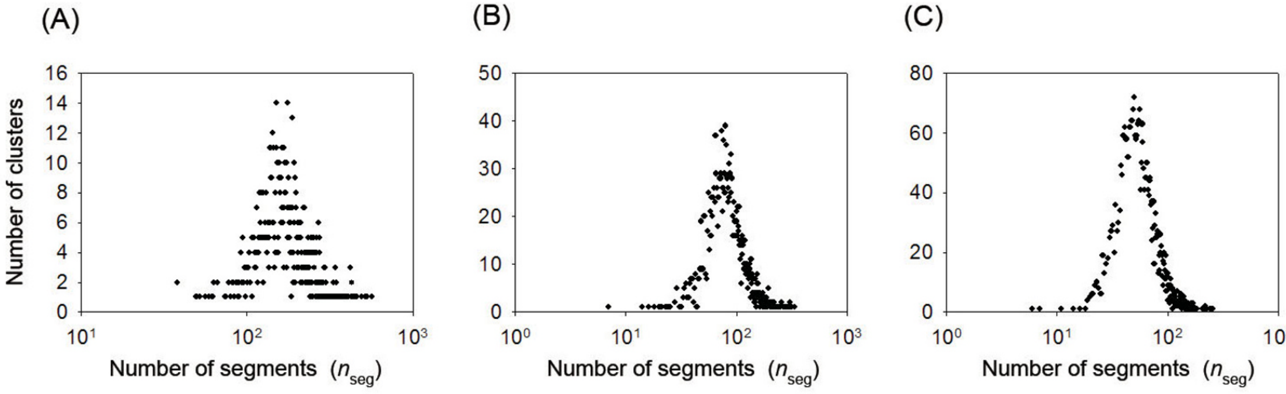


Figure 8
Relation between number (n_{seg}) of segments involved in a cluster and number of clusters for K_c = 1000 (A), 2000 (B), and 3000 (C).

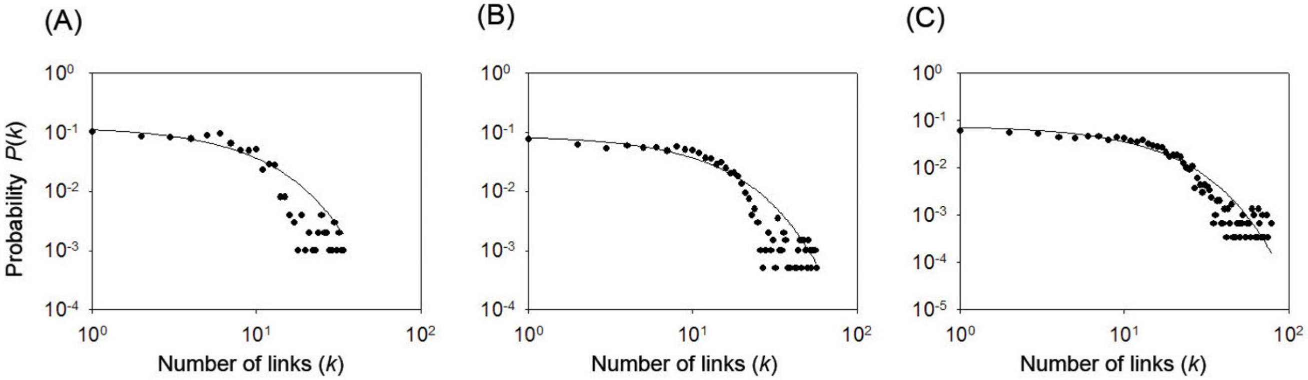


Figure 9
Connectivity distribution $P(k)$ of cluster network at K_c = 1000 (A), 2000 (B), and 3000 (C). The X-axis k shows the number of links of a cluster connected to other clusters. Solid lines are the best-fit curves drawn assuming that $P(k)$ decays with k exponentially.

We next calculated the number of communities at various K_c. We classified the communities into major and minor communities. Major ones are communities consisting of more than three clusters. Then, minor ones are small isolated communities consisting of only one or two clusters without links to other communities. No community involves only one cluster linked to another community. The K_c dependence of the number (N_{com}) of the major communities is presented in Figure 10B. The minor communities do not characterize the overall property of the network because only 10% of clusters belong to the minor communities at any K_c. The increment of N_{com} with increasing K_c was rapid for $100 \leq K_c \leq 1000$ and slow for $K_c \geq 1000$. The values of N_{com} were, respectively, 36, 38, and 38 at K_c = 1000, 2000, and 3000. This result shows that the number of communities was conserved for $K_c \geq 1000$.

In addition to the analysis presented above, we checked to determine whether the contents (i.e., segments) involved in the communities are conserved with variation of K_c. Subsequently, we assigned a single color to communities common to the universes at K_c = 1000 (Figure 11A), 2000 (Figure 11B), and 3000 (Figure 11C). For instance, the majority of segments in the orange community of Figure 11A were involved in the orange ones in Figures 11B and

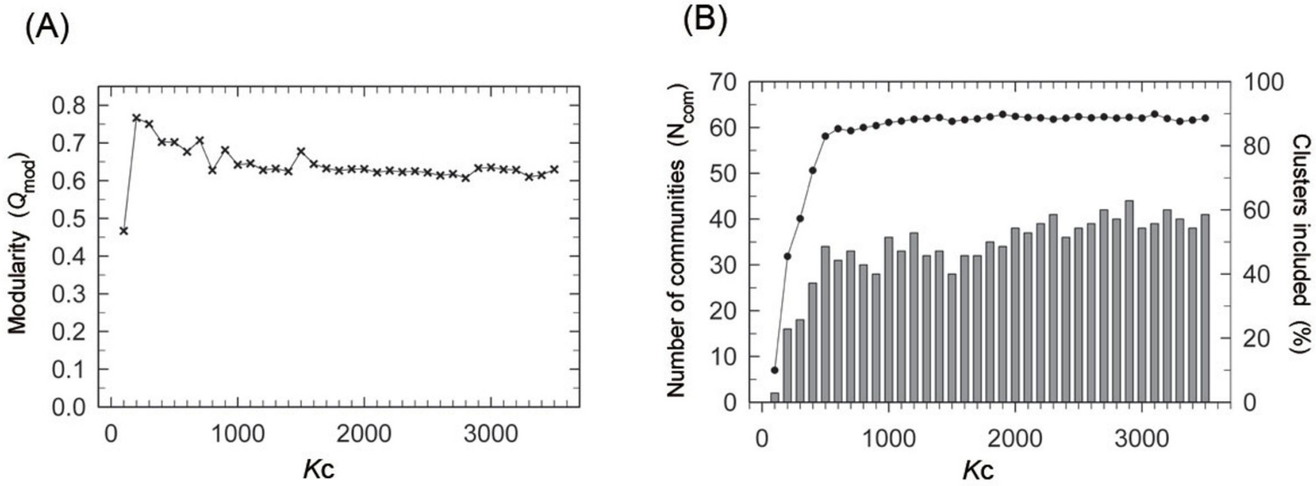


Figure 10
K_c dependence of N_{com} and Q_{mod}. (A) The K_c dependence of modularity Q_{mod} (Eq. 10). (B) The bar graph shows the K_c dependence of number, N_{com}, of communities assigned to the left y-axis. The line with filled circles represents the ratio (assigned to right y-axis) of clusters in major communities to all clusters.

11C. Consequently, the communities are conserved well in the universes at different K_c. In other words, the network partitioning into communities is universal, independent of the spatial resolution (i.e., K_c). We termed this inter-community partitioning as "sub-partitioning", whereas the main partitioning is inter-regional partitioning (Figure 5).

Discussion

Herein, we described universal partitioning of two types in the 50-residue segment networks (Figure 5) based on the network analysis. The main partitioning (the network separation by fold regions) resembles that in the classification scheme of existing databases such as CATH and SCOP. The mainly α, mainly β, αβ, and randomly structured regions consist respectively of α, β, αβ, and randomly structured communities. However, for the first time, we found communities in the segment fold universe: this sub-partitioning (network separation by communities) is a novel finding. High modularity ensures persistently existing communities, where the intra-community clusters are linked tightly and the inter-community clusters are linked weakly. The universality of the sub-partitioning was remarkable for f_0 ($0.65 \leq f_0 \leq 0.75$). Nevertheless, outside this range, the universality vanishes gradually. Our results reveal a hierarchically structured universe for 50-residue segments, as depicted in Figure 12. This hierarchy is robust because the main and sub-partitionings are independent of K_c for $K_c \geq 1000$.

Figure 10B portrays that the current universe for the 50-residue segments consists of some dozens (ca. 40) of major communities. Kihara and Skolnick reported that the current PDB database might cover almost all structures of small proteins [50]. Crippen and Maiorov generated many self-avoiding conformations of a chain and suggested that the possible structures of a 50-residue chain are classifiable roughly into a small number of types, although the secondary-structure formation was not incorporated in their model [51]. A study proposed the conjecture that tertiary-structure evolution of proteins might be achieved using limited repertoires of basic units such as supersecondary structure elements [52]. Results of such studies are consistent with our results because we have shown that protein tertiary structures can be decomposed into the dozens of major communities of 50-residue segments. Actually, 90% of clusters belong to the major communities. To link those studies with our study more closely, detailed contents of each major community should be investigated. In fact, such a research project is proceeding now. Moreover, the role of the minor communities in the protein structure construction should be studied.

The currently observed 50-residue segment universe was characterized by the non-power-law distribution. Our result apparently differs from the power-law distribution widely known for the hierarchical protein domain universe [9,46,47,53]. The emergence of the non-power-law statistics might be related to the usage of the inter-residue contact, which is a more relaxed similarity measure than widely used ones such as RMSD or the DALI Z-score. It is known that in the power-law statistics the rate for isolated clusters in the entire clusters is high [53]. In our non-power law statistics, the rate was low because the relaxed measure provided linkages between clusters. Thus, the two statistics compensate to each other to survey the fold universe. From the non-power-law universe, we could show a novel hierarchy (Figure 12) in the universe and the existence of 40 repertories (Figure 10) to construct the protein tertiary structures, which have not been reported from the power-law universe. These results were also

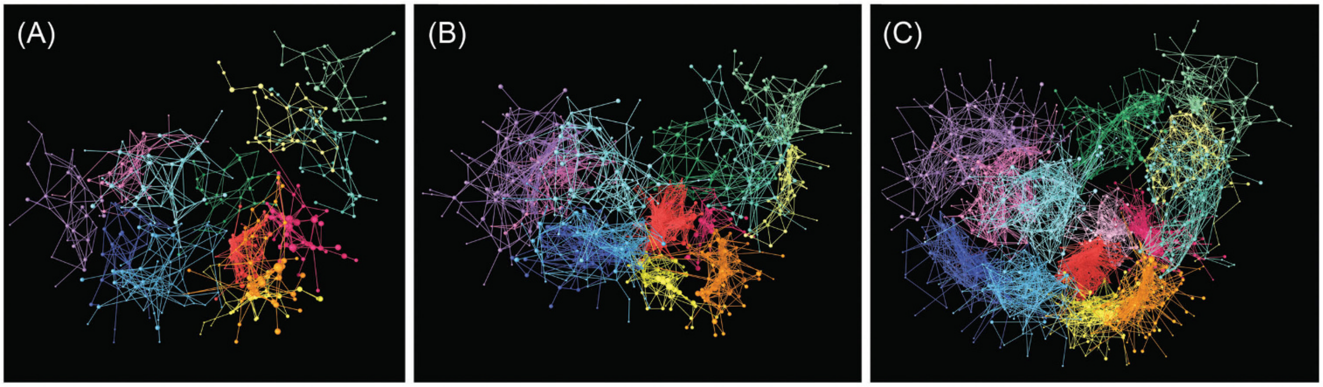


Figure 11
Communities at K_c = 1000 (A), 2000 (B), and 3000 (C). For each universe, only the top 13 communities by the number of involved clusters are shown. A single color is assigned to communities that are common to the three universes. Communities that are not common among the three are not shown, nor are minor communities.

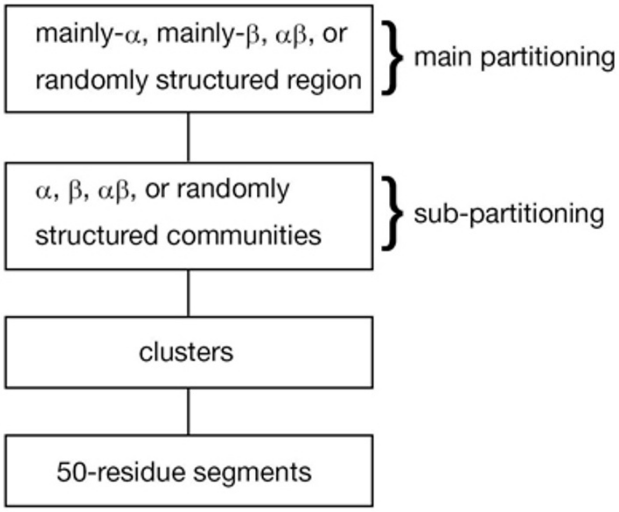
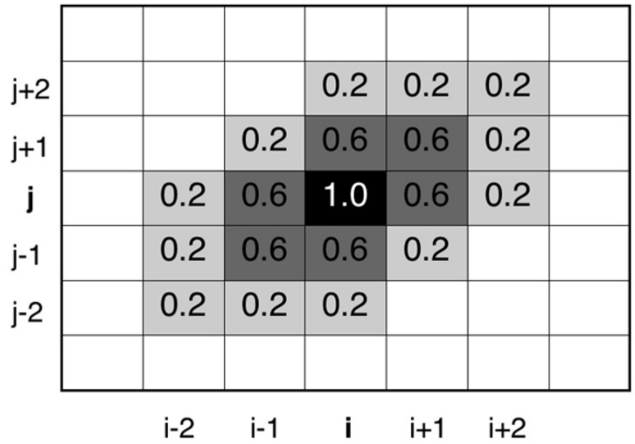

mainly-α, mainly-β, αβ, or randomly structured region
main partitioning
α, β, αβ, or randomly structured communities
sub-partitioning
clusters
50-residue segments

Figure 12
Hierarchy in the segment universe proposed from the current study.

found in the 60- and 70-residue segment universes (data not shown). This suggests that the non-power law is likely to be a general property for segment universes.

The current network helps to trace conformational changes of segments along the network linkages. *Supplementary Results* displays that the conformation gradually changes when shifting the view from a cluster to another (see Additional file 1).



j+2			0.2	0.2	0.2
j+1		0.2	0.6	0.6	0.2
j	0.2	0.6	1.0	0.6	0.2
j-1	0.2	0.6	0.6	0.2	
j-2	0.2	0.2	0.2		
	i-2	i-1	i	i+1	i+2

Figure 13
Smoothed inter-residue contacts *c*(*i, j*) (Eq. 4). It is presumed that residue pair (*i*, *j*) is in contact (i.e., $c(i, j) = 1$), and that the other pairs are non-contacting. Equation 4 provides negative $c_s(i', j')$ at sites where an inequality, $|i - i'| + |j - j'| + |(|i - i'| - |j - j'|)| > 5$, is satisfied. Besides, this inequality is satisfied without exception when any one of the three inequalities, $|i - i'| > 2$, $|j - j'| > 2$, or $||i - i'| - |j - j'|| > 2$, is met. Those negative $c(i, j) = 1$), and that the other pairs are non-contacting. Equation 4 provides negative $c_s(i', j')$ are reset to zero (see text).

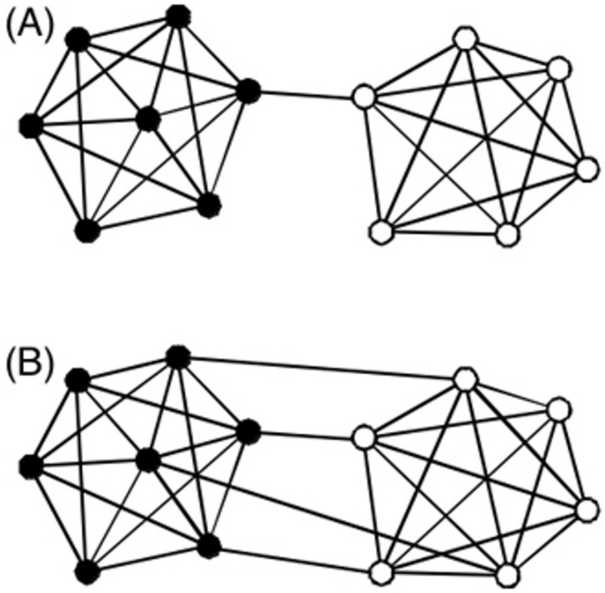


Figure 14
Two network types. Network (A) has larger modularity Q_{mod} than (B) does. Filled circles form a community (Com 1); open ones construct the other community (Com 2). Lines between circles represent links.

The inter-residue contact (native contact) has been widely used as a reaction coordinate in protein folding (see *Introduction*). We intend to use the currently obtained networks for protein folding study. The networks of fixed-length segments are readily applicable for conformational sampling in protein folding, where the chain length is usually fixed. The randomly structured clusters are located at the root of the distribution (Figure 4 and Figure 5), from which the segment conformation can diversify to mainly α, mainly β, or αβ regions with increased compactness (Figure 7).

Conclusion

We constructed a 50-residue segment network for investigating the protein local structure universe. The network was partitioned into some dozens (ca. 40) of major communities with high modularity ($0.60 < Q_{mod} < 0.65$), independent of the spatial resolution (K_c). The major communities existed universally and persistently in the universe. Surprisingly, 90% of all segments were covered by the major communities. Consequently, numerous similarities exist among local regions (i.e., 50-residue segments) of proteins. Furthermore, the currently constructed segments networks are characterized by non-power-law (non-scale-free) statistics, which apparently differs from reported characteristics for the fold universe of full-length proteins.

Methods

This section includes six subsections. The first three – "Generation of 50-residue segment library", "Clustering segments", and "Computation of inter-residue contact patterns" – are preparative subsections describing construction of the 50-residue segment fold universe. In the subsection titled "Construction of a universe and network", construction of the fold universe and the network is described. "Modularity analysis" presents analyses used to examine the network. The subsection "Characterization of communities by structural features" describes a method to characterize communities depending on five structural features. Specification of indices *i*, *j*, *s*, *t*, *u*, *v*, and *w* is given at the beginning of *Results*.

Generation of 50-residue segment library

We generated a structure library of 50-residue segments with reference to the all-α, all-β, α/β, and α+β fold classes defined in the SCOP database (release 1.69) [5]. The SCOP database presents a list that provides a representative for each protein family. We selected tertiary structures of the representative domains from the PDB database [54] with elimination of multi-chain domains, those involving structurally undetermined regions, and those shorter than 50 residues. Furthermore, we eliminated domains consisting of 400 residues or more, which might involve structurally repeating units. Then we obtained 1803 domains (456 from all-α, 393 from all-β, 393 from α/β, and 561 from α+β). A domain that is n_r amino-acid residues long produces n_r - 49 segments from sliding a 50-residue window along the sequence one residue-by-one residue. Finally, we obtained an ensemble of 186 821 segments (32 040 from all-α, 39 375 from all-β, 63 177 from α/β, and 52 229 from α+β). The residue site of each segment was re-numbered from 1 to 50 in our study.

Clustering segments

We classify the collected segments into clusters as follows: First, the inter-C_α atomic distances were calculated for segment *s*, where the distance between residues *i* and *j* is denoted as $r_s(i, j)$. We eliminated residue pairs $|i - j| < 3$ because the distances for these pairs are similar for all segments. In other words, those distances have less sensitivity to discriminate the structural differences of segments. Then, the number (N_{pair}) of the C_α-atomic pairs in a 50-residue segment is 1128: $N_{\text{pair}} = 1128$. The set of distances is expressed as a N_{pair}-dimensional vector: $\vec{r}_s = [r_s(1, 4), r_s(1, 5), ..., r_s(47, 50)]$. We define the root mean square distance ($rmsd_{st}$) between $\vec{r}_s$ and $\vec{r}_t$ as in the N_{pair}-dimensional Cartesian space: $rmsd_{st} = |\vec{r}_s - \vec{r}_t|$.

For classifying the 186 821 segments into K_c clusters, we applied Lloyd's K-means algorithm [55] to the set of $rmsd_{st}$ values, where *s*, *t* = 1, ..., 186821. One should set K_c in advance in the K-means algorithm. We examined various values for K_c ($K_c \leq 5000$). In Lloyd's method, the K_c clusters are set randomly at the beginning. The finally converged clusters are output. We have checked that the main results are independent of the initial set of clusters.

We calculated the center ($\bar{u}$) of a cluster *u* in the N_{pair}-dimensional space as $\vec{r}_{\bar{u}} = [r_{\bar{u}}(1,4), r_{\bar{u}}(1,5), ..., r_{\bar{u}}(47,50)]$, where the element $r_{\bar{u}}(i, j)$ is given as

$$r_{\bar{u}}(i,j) = \frac{\sum_{s \in u} r_s(i,j)}{n_u}. \quad (1)$$

The n_u is the number of constituent segments of the cluster *u*.

We defined a size S_u of the cluster *u* as

$$S_u = \frac{\sum_{s \in u} rmsd_{s\bar{u}}}{n_u}. \quad (2)$$

This equation simply quantifies the average distance from the cluster center $\bar{u}$ to segments belonging to the cluster *u* in the N_{pair}-dimensional space. The average cluster size is defined simply as

$$< S >= \frac{\sum_{u}^{K_c} S_u}{K_c}, \quad (3)$$

where the summation is taken over all the K_c clusters.

Computation of inter-residue contact patterns

In this subsection, we present computation of the inter-cluster and intra-cluster structural similarity based on the inter-residue contact patterns. The inter-residue contacts in segment *s* were defined as follows: Calculating all the inter-heavy atomic distances between residues *i* and *j* for the segment, their minimum distance was registered as the inter-residue distance $q_s(i, j)$. Then, if $q_s(i, j) < 6.0$ Å, we judged that the residues *i* and *j* were contacting and set a quantity $c_s(i, j)$ to 1 (otherwise, $c_s(i, j) = 0$). Here, we again eliminated residue pairs of $|i - j| < 3$ in the calculation of

$c_s(i, j)$. The set of $c_s(i, j)$ constructs a matrix C_s, where element (i, j) is $c_s(i, j)$.

The upper limit (6.0 Å) for $q_s(i, j)$ allows no penetration of a water molecule between residues *i* and *j*: At $q_s(i, j) = 6.0$ Å, the substantial space for water penetration between the residues is approximately 2.0 Å (= 6.0 - 2 × 2.0) assuming that radii of segment heavy atoms are 2.0 Å. This space of 2.0 Å is smaller than the diameter of a water molecule (2.8 Å).

A structural similarity between segments *s* and *t* might be counted by comparing C_s and C_t. However, a strict comparison engenders an oversight of the similarity in the following case: Presume that $c_s(i, j) = 1$ and $c_t(i,+ 1, j) = 0$ in the segment *s*, and $c_s(i, j) = 0$ and $c_t(i,+ 1, j) = 1$ in segment *t*. The inter-residue contacts in these segments differ but they are similar. The strict comparison does not count such a similarity. To incorporate such similarity, smoothing of C_s was performed as

$$c_s(i', j') = 1.0 - 0.2[|\, i - i' \,| + |\, j - j' \,| + |\, (|\, i - j \,| - |\, i' - j' \,|) \,|]. \tag{4}$$

This smoothing (see Figure 13) was done only when residues *i*' and *j*' are not contacting and the residues *i* and *j* are contacting in the segment. If Eq. 4 produces a negative value, then $c_s(i', j')$ is set to zero. If a non-contacting residue pair (i', j') has multiple values for $c_s(i', j')$ attributable to contributions of some contacting pairs around (i', j'), then the largest value is assigned to the non-contacting pair. As described in this paper, the inter-residue contact matrix C_s indicates that after the smoothing.

Here, we calculate the contact patterns which are specific to a cluster. For this purpose, we averaged *C* over the entire segment library and over all segments in cluster *u*. We denote these averaged matrices as $\bar{C}_{whole}$ and $\bar{C}_u$, respectively. Then, we defined a quantity $\Delta\bar{C}_u = \bar{C}_u - \bar{C}_{whole}$, where element (i, j) is denoted as $\Delta\bar{c}_u(i, j)$. The similarity between clusters *u* and *v* was measured using the following correlation coefficient:

$$f(\Delta\bar{C}_u, \Delta\bar{C}_v) = \frac{<\Delta\bar{C}_u\Delta\bar{C}_v> \;\; <\Delta\bar{C}_u><\Delta\bar{C}_v>}{[<\Delta\bar{C}_u^2>-<\Delta\bar{C}_u>^2]^{1/2}[<\Delta\bar{C}_v^2>-<\Delta\bar{C}_v>^2]^{1/2}} \tag{5}$$

where

$$<\Delta\bar{C}_u\Delta\bar{C}_v> = \frac{\sum_{i,j}^{N_{\text{pair}}} \Delta\bar{c}_u(i,j)\Delta\bar{c}_v(i,j)}{N_{\text{pair}}}. \tag{6}$$

The term $<\Delta\bar{C}_u^2>$ in Eq. 5 is defined by setting $u = v$ in Eq. 6, and the term $<\Delta\bar{C}_u>$ by setting $\Delta\bar{C}_v = 1$. A large correlation coefficient indicates similar inter-residue contact patterns between the clusters.

The coefficient $f(\Delta\bar{C}_u, \Delta\bar{C}_v)$ is useful as a distance between clusters *u* and *v* in a multi dimensional space. Consequently, the set of coefficients define a multidimensional weighted graph (i.e., weighted network). In this work, we must convert this weighted graph into an un-weighted one to perform community analysis, which only deals with the un-weighted graph. Therefore, we introduce an adjacency matrix a_{uv} in which element (u, v) is given as follows.

$$\begin{aligned} a_{uv} &= 1 \quad (\text{for } f(\Delta\bar{C}_u, \Delta\bar{C}_v) > f_0) \\ &= 0 \quad (\text{otherwise}) \end{aligned} \tag{7}$$

The inter-residue contact patterns are similar between clusters *u* and *v* only when $f(\Delta\bar{C}_u, \Delta\bar{C}_v) > f_0$. Herein, we set f_0 to 0.7. The meaning of 0.7 is explained in the *Results* section.

We next assessed the intra-cluster similarity. First, we defined a quantity $\Delta C_s = C_s - \bar{C}_{whole}$ for a segment *s*, where element (i, j) of ΔC_s is denoted as $\Delta C_s(i, j)$. Then, we averaged $\Delta C_s(i, j)$ over the segments in cluster *u*:

$$g_u(i, j) = \frac{\sum_{s \in u} \Delta c_s(i,j)}{n_u}. \tag{8}$$

We define a matrix G_u for that the element (i, j) as $g_u(i, j)$. Then, we calculated the correlation coefficient $f(G_u, \Delta C_s)$ between G_u and ΔC_s for segments in cluster *u*, using the same definition as that in Eq. 5. Subsequently, we calculated an averaged correlation coefficient $<f>_u$ over $f(G_u,\Delta C_s)$ of the segments in the cluster *u*. This quantity is a measure to express the similarity of the inter-residue contact patterns among the segments in cluster *u*. Finally, $<f>_u$ was averaged over all clusters.

$$<f>_{K_c} = \frac{\sum_{u=1}^{K_c} <f>_u}{K_c} \tag{9}$$

The larger the value of $<f>_{K_c}$, the more similar the inter-residue contact patterns in each cluster are, on average.

Construction of a universe and network

We constructed a distribution (i.e., fold universe) of K_c clusters in a 3D conformational space with embedding clusters into the 3D. Details are presented in Additional file 1. As explained in the *Introduction*, lowering of the space dimensionality hides the internal architecture of the fold universe. To compensate the full-dimensional information to the 3D distribution, links were assigned to clusters with similar inter-residue contact patterns ($a_{uv} = 1$). The generated networks were subjected to the modularity analysis described in the next subsection.

Modularity analysis

To investigate a property of the cluster network, we divided the network into communities (i.e., sub-networks) using an efficient method [44]. An example of a network is presented in Figure 14, where two communities (Com 1 and Com 2) exist. A modularity Q_{mod} is an index to assess how well the network is divided into communities [49]:

$$Q_{\mathrm{mod}} = \sum_{w=1}^{N_{\mathrm{com}}} [I_w / I - (d_w / 2I)^2], \tag{10}$$

where I_w is the number of links connecting clusters within a community w, N_{com} is the number of communities existing in the entire network, and I is the number of links existing in the entire network. The quantity d_w is called the "total degree", which is defined for each community as $d_w = 2I_{\mathrm{w}} + I_{\mathrm{w\text{-}other}}$, where $I_{\mathrm{w\text{-}other}}$ is the number of links connecting clusters in the community w and clusters outside the community. The value of Q_{mod} is 0–1: Q_{mod} approaches 1 when the number of links connecting different communities decreases. For instance, the network in Figure 14A has Q_{mod} of 0.466 ($I = 34$, $I_1 = 18$, $I_2 = 15$, $d_1 = 37$, and $d_2 = 31$). That of Figure 14B has Q_{mod} of 0.388 ($I = 37$, $I_1 = 18$, $I_2 = 15$, $d_1 = 40$, and $d_2 = 34$). The two networks are equivalent except for the inter-community links.

Characterization of communities by structural features

The manner of differentiating the communities is important. Herein, we characterize the communities depending on five biophysical structural features: radius of gyration (R_{g}), number of inter-residue contacts ($N_{\mathrm{contact}} = \sum_{i,j}^{N_{\mathrm{pair}}} c(i, j)$ with removal of pairs of $|i - j| < 3$), number of α-helical residues (n_α), number of β-helical residues (n_β), and the sum of n_α and n_β (i.e., $n_{\alpha\beta} = n_\alpha + n_\beta$).

First, we calculate the five quantities for each segment. The secondary-structure assignment to each residue in a segment is done using software available at the STRIDE web site http://webclu.bio.wzw.tum.de/stride/[56]. Next, we took the average for each of the five quantities over segments in a community. We designate the average quantities in a community w as $R_{\mathrm{g}}(w)$, $N_{\mathrm{contact}}(w)$, $n_\alpha(w)$, $n_\beta(w)$, and $n_{\alpha\beta}(w)$. Then, we classify the communities into α, β, αβ, and randomly structured ones according to the five quantities: Randomly structured communities are those with $R_{\mathrm{g}} > 14$ Å and $N_{\mathrm{contact}}(w) < 100$ or those with $n_{\alpha\beta}(w) < 15$. In the remaining communities, α communities are those with $n_\alpha(w) > 0.7 \times n_{\alpha\beta}(w)$. In the remaining communities, β communities are those with $n_\alpha(w) > 0.7 \times n_{\alpha\beta}(w)$. The finally remaining communities are classified as αβ communities. Each segment in the αβ communities significantly involves both an α helix and a β strand.

Authors' contributions

This study was conceived and carried out by JI, who also developed the main part of the methodology. YS participated in some analyses. IK participated in discussions. KT participated in the coordination of the study. He also helped to write the manuscript. JH participated in developing the methodology, designed the study, and wrote the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Supplementary Methods and Supplementary Results*. There are three sections in the Supplementary Methods as follows: (1) The method of embedding the inter-cluster network into 3D space. (2) The definition of F-measure. (3) The coloring method for clusters in the 3D network. In the Supplementary Results, tertiary structures of fragments in the same cluster and those in the same community are discussed.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1472-6807-9-34-S1.doc]

Acknowledgements

KI and JH were partly supported by **BIRD** of Japan Science and Technology Agency (**JST**). JH was also partly supported by New Energy and Industrial Technology Development Organization (**NEDO**).

References

1. Chothia C: **Proteins. One thousand families for the molecular biologist.** *Nature* 1992, **357:**543-544.
2. Gibrat JF, Madej T, Bryant SH: **Surprising similarities in structure comparison.** *Curr Opin Struct Biol* 1996, **6:**377-385.
3. Coulson AFW, Moult J: **A unifold, mesofold, and superfold model of protein fold use.** *Proteins* 2002, **46:**61-71.
4. Liu X, Fan K, Wang W: **The number of protein folds and their distribution over families in nature.** *Proteins* 2004, **54:**491-499.
5. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247:**536-540.
6. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH – a hierarchic classification of protein domain structures.** *Structure* 1997, **5:**1093-1108.

7. Efimov AV: **Structural trees for protein superfamilies.** *Proteins* 1997, **28:**241-260.
8. Holm L, Sander C: **Mapping the protein universe.** *Science* 1996, **273:**595-602.
9. Dokholyan NV, Shakhnovich B, Shakhnovich EI: **Expanding protein universe and its origin from the biological Big Bang.** *Proc Natl Acad Sci USA* 2002, **99:**14132-14136.
10. Hou J, Sims GE, Zhang C, Kim S-H: **A global representation of the protein fold space.** *Proc Natl Acad Sci USA* 2003, **100:**2386-2390.
11. Hou J, Jun S-R, Zhang C, Kim S-H: **Global mapping of the protein structure space and application in structure-based inference of protein function.** *Proc Natl Acad Sci USA* 2005, **102:**3651-3656.
12. Holm L, Sander C: **Protein structure comparison by alignment of distance matrices.** *J Mol Biol* 1993, **233:**123-138.
13. Orengo CA, Flores TP, Taylor WR, Thornton JM: **Identification and classification of protein fold families.** *Protein Eng* 1993, **6:**485-500.
14. Standley DM, Kinjo AR, Kinoshita K, Nakamura H: **Protein structure databases with new web services for structural biology and biomedical research.** *Brief Bioinfo* 2008, **9:**276-285.
15. Takahashi K, Go N: **Conformational classification of short backbone fragments in globular proteins and its use for coding backbone conformations.** *Biophys Chem* 1993, **47:**163-178.
16. Tomii K, Kanehisa M: **Systematic detection of protein structural motifs.** In *Pattern discovery in biomolecular data* Edited by: Wang JTL, Shapiro BA, Shasha D. New York: Oxford University Press; 1999:97-110.
17. Choi IG, Kwon J, Kim S-H: **Local feature frequency profile: A method to measure structural similarity in proteins.** *Proc Natl Acad Sci USA* 2004, **101:**3797-3802.
18. Ikeda K, Tomii K, Yokomizo T, Mitomo D, Maruyama K, Suzuki S, Higo J: **Visualization of conformational distribution of short to medium size segments in globular proteins and identification of local structural motifs.** *Protein Sci* 2005, **14:**1253-1265.
19. Sawada Y, Honda S: **Structural diversity of protein segments follows a power-law distribution.** *Biophys J* 2006, **91:**1213-1223.
20. Ikeda K, Hirokawa T, Higo H, Tomii K: **Protein-segment universe exhibiting transitions at intermediate segment length in conformational subspaces.** *BMC Structural Biology* 2008, **8:**37.
21. Simons KT, Kooperberg C, Huang E, Baker D: **Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.** *J Mol Biol* 1997, **268:**209-225.
22. Bonneau R, Strauss CE, Rohl CA, Chivian D, Bradley P, Malmström L, Robertson T, Baker D: **De novo prediction of three-dimensional structures for major protein families.** *J Mol Biol* 2002, **322:**65-78.
23. Chikenji G, Fujitsuka Y, Takada S: **A reversible fragment assembly method for de novo protein structure prediction.** *J Chem Phys* 2003, **119:**6895-6903.
24. Jeong H, Mason SP, Barabási AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411:**41-42.
25. Holme P, Huss M, Jeong H: **Subnetwork hierarchies of biochemical pathways.** *Bioinformatics* 2003, **19:**532-538.
26. Guimerà R, Amaral LAN: **Functional cartography of complex metabolic networks.** *Nature* 2005, **433:**895-900.
27. Palla G, Derényi I, Farkas I, Vicsek T: **Uncovering the overlapping community structure of complex net-works in nature and society.** *Nature* 2005, **435:**814-818.
28. Go N: **Theoretical studies of protein folding.** *Annu Rev Biophys Bioeng* 1983, **12:**183-210.
29. Go N, Abe H: **Randomness of the process of protein folding.** *Int J Pept Protein Res* 1983, **22:**622-632.
30. Wolynes PG, Onuchic JN, Thirumalai D: **Navigating the folding routes.** *Science* 1995, **267:**1619-1620.
31. Galzitskaya OV, Finkelstein AV: **A theoretical search for folding/unfolding nuclei in three-dimensional protein structures.** *Proc Natl Acad Sci USA* 1999, **96:**11229-11304.
32. Munoz V, Eaton WA: **A simple model for calculating the kinetics of protein folding from three-dimensional structures.** *Proc Natl Acad Sci USA* 1999, **96:**11311-11316.
33. Shea J-E, Brooks CL III: **From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding.** *Annu Rev Phys Chem* 2001, **52:**499-535.
34. Koga N, Takada S: **Roles of native topology and chain-length scaling in protein folding: A simulation study with a Go-like model.** *J Mol Biol* 2001, **313:**171-180.
35. Makarov DE, Keller CA, Plaxco KW, Metiu H: **How the folding rate constant of simple, single-domain proteins depends on the number of native contacts.** *Porc Natl Acad Sci USA* 2002, **99:**3535-3539.
36. Zhou HX: **Theory for the rate of contact formation in a polymer chain with local conformational transitions.** *J Chem Phys* 2003, **118:**2010-2015.
37. Nakamura HK, Sasai M, Takano M: **Scrutinizing the squeezed exponential kinetics observed in the folding simulation of an off-lattice Go-like protein model.** *Chem Phys* 2004, **307:**259-267.
38. Mitomo D, Nakamura HK, Ikeda K, Yamagishi A, Higo J: **Transition state of a SH3 domain detected with principle component analysis and a charge-neutralized all-atom protein model.** *Proteins* 2006, **64:**883-894.
39. Ikebe J, Kamiya N, Shindo H, Nakamura H, Higo J: **Conformational sampling of a 40-residue protein consisting of α and β secondary-structure elements in explicit solvent.** *Chem Phys Lett* 2007, **443:**364-368.
40. Kamiya N, Mitomo D, Shea J-E, Higo J: **Folding of the 25 residue Abeta(12–36) peptide in TFE/water: temperature-dependent transition from a funneled free-energy landscape to a rugged one.** *J Phys Chem B* 2007, **111:**5351-5356.
41. Baker D: **A surprising simplicity to protein folding.** *Nature* 2000, **405:**39-42.
42. Kamagata K, Arai M, Kuwajima K: **Unification of the folding mechanisms of non-two-state and two-state proteins.** *J Mol Biol* 2004, **339:**951-965.
43. Kamagata K, Kuwajima K: **Surprisingly high correlation between early and late stages in non-two-state protein folding.** *J Mol Biol* 2006, **357:**1647-1654.
44. Newman MEJ: **Finding community structure in net-works using the eigenvectors of matrices.** *Phys Rev E* 2006, **74:**036104.
45. Grant A, Lee D, Orengo C: **Progress towards mapping the universe of protein folds.** *GenomeBiology* 2004, **5:**107.
46. Koonin EV, Wolf YI, Karev GP: **The structure of the protein universe and genome evolution.** *Nature* 2002, **420:**218-223.
47. Qian J, Luscombe NM, Gerstein M: **Protein Family and Fold Occurrence in Genomes: Power-law Behaviour and Evolutionary Model.** *J Mol Biol* 2001, **313:**673-681.
48. Barabási AL, Albert R: **Emergence of scaling in random networks.** *Science* 1999, **286:**509-512.
49. Newman MEJ, Girvan M: **Fast algorithm for detecting community structure in networks.** *Phys Rev E* 2004, **69:**026113.
50. Kihara D, Skolnick J: **The PDB is a covering set of small protein structures.** *J Mol Biol* 2003, **334:**793-802.
51. Crippen GM, Maiorov VN: **How Many Protein Folding Motifs are There?** *J Mol Biol* 1995, **252:**144-151.
52. Soding J, Lupas AN: **More than the sum of their parts: on the evolution of proteins from peptides.** *BioEssay* 2003, **25:**837-846.
53. Krishnadev O, Brinda KV, Vishveshwara S: **A graph spectral analysis of the structural similarity of protein chains.** *Proteins* 2005, **61:**152-163.
54. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The protein data bank.** *Nucleic Acids Res* 2000, **28:**235-242.
55. Lloyd SP: **Least squares quantization in PCM.** *IEEE Transactions on Information Theory* 1982, **28:**129-137.
56. Frishman D, Argos P: **Knowledge-based protein secondary structure assignment.** *Proteins* 1995, **23:**566-579.