Journal of Theoretical Biology 258 (2009) 294–301

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Donut-shaped fingerprint in homologous polypeptide relationships—A topological feature related to pathogenic structural changes in conformational disease

Xin Liu [a], Ya-Pu Zhao [b,*]

[a] Institute of Mechanics, Chinese Academy of Sciences, Beijing 100080, China
[b] The State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences. Beijing 100190, China

ARTICLE INFO

Article history:
Received 19 September 2008
Received in revised form
6 January 2009
Accepted 11 February 2009
Available online 25 February 2009

Keywords:
Homologous relationship
Polypeptide
Prion protein
Conformational disease

ABSTRACT

Features of homologous relationship of proteins can provide us a general picture of protein universe, assist protein design and analysis, and further our comprehension of the evolution of organisms. Here we carried out a study of the evolution of protein molecules by investigating homologous relationships among residue segments. The motive was to identify detailed topological features of homologous relationships for short residue segments in the whole protein universe. Based on the data of a large number of non-redundant proteins, the universe of non-membrane polypeptide was analyzed by considering both residue mutations and structural conservation. By connecting homologous segments with edges, we obtained a homologous relationship network of the whole universe of short residue segments, which we named the graph of polypeptide relationships (GPR). Since the network is extremely complicated for topological transitions, to obtain an in-depth understanding, only subgraphs composed of vital nodes of the GPR were analyzed. Such analysis of vital subgraphs of the GPR revealed a donut-shaped fingerprint. Utilization of this topological feature revealed the switch sites (where the beginning of exposure of previously hidden "hot spots" of fibril-forming happens, in consequence a further opportunity for protein aggregation is provided; 188–202) of the conformational conversion of the normal α-helix-rich prion protein PrP^{C} to the β-sheet-rich PrP^{Sc} that is thought to be responsible for a group of fatal neurodegenerative diseases, transmissible spongiform encephalopathies. Efforts in analyzing other proteins related to various conformational diseases are also introduced.

1. Background

Computational approaches, such as homology modeling (Chou, 2004a), structural bioinformatics (Chou, 2004b; Liu et al., 2008b), pharmacophore modeling (Sirois et al., 2004), Monte Carlo simulated annealing (Chou, 1992), protein subcellular location prediction (Chou and Shen, 2007b, 2008), and signal peptide prediction (Chou and Shen, 2007a; Shen and Chou, 2007), can provide very useful information on and insight into basic research and drug design. Since our ability to characterize the biological properties of a protein is almost exclusively based on properties conserved through evolutionary time, the study of protein evolution using computational approaches has been the focus of many researchers (Socolich et al., 2005; Russ et al., 2005; Zhang and Liu, 2008; Liu et al., 2008a).

Characterization of the protein universe can assist in comprehending the evolvement, i.e., the formation, past, and future of proteins, in designing artificial proteins, and in providing information useful in other biological fields. A well-known feature of the protein universe is that some folds are abundantly represented by proteins with sequence identity as low as random sequences (Rost, 1997; Holm and Sander, 1993, 1997), whereas other folds are represented by a single sequence (Teichmann et al., 1999; Orengo et al., 1999; Holm and Sander, 1996). To explain this variability in fold representation, it has been suggested that a premise in convergent evolution is that folds with higher designability can be encoded and represented by more sequences (Finkelstein et al., 1995; Govindarajan and Goldstein, 1996; Li et al., 1996). This phenomenological notion was identified from the observation of exhaustive sequence enumeration in a lattice protein model. Application of this notion led to remarkable progress in research in areas such as folding mechanisms (Li et al., 1998; Wolynes, 1996; England et al., 2003), plotting of the distribution of protein populations (Taverna and Goldstein, 2000; Shakhnovich et al., 2005), and hereditary diseases (Wong and Frishman, 2006). However, the designability principle often provides features of a lattice model, but not of actual proteins. Several attempts have been made to define a more realistic picture of homologous relationships.

* Corresponding author Te.: +86 10 82543932; fax: +86 10 82543977.
E-mail addresses: liuxin@lnm.imech.ac.cn (X. Liu), yzhao@imech.ac.cn (Y.-P. Zhao).

0022-5193/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2009.02.009

Dokholyan et al. (2002) offered a general picture of the universe of protein structure. Based on structural alignments provided by the FSSP database, the authors claimed that the graph formed by proteins/vertices of a non-redundant set and connections/edges between any two structurally similar protein domains was a scale-free network. In such a network, the probability density $P(K)$ of a domain with K related structures (connection number) follows a power law $P(K) = K^{-\alpha}$. Several similar studies have been performed based on sequence, structure, or both (Huynen and van Nimwegen, 1998; Yanai et al., 2000; Qian et al., 2001; Koonin et al., 2002). All the networks obtained have the same scale-free feature. In fact, as a feasible method to obtain useful insights, graphical approaches have been used in the study of many biological systems, such as enzyme-catalyzed reactions (Andraos, 2008; Chou, 1989; Chou and Forsen, 1980; Kuzmic et al., 1992; Myers and Palmer, 1985; Zhou and Deng, 1984), protein folding kinetics (Chou, 1990), inhibition kinetics of processive nucleic acid polymerases and nucleases (Althaus et al., 1993; Chou and Kezdy, 1994), analysis of codon usage (Chou and Zhang, 1992; Zhang and Chou, 1994), analysis of DNA sequences (Qi et al., 2007), among others. Moreover, in the recent years, graphical methods have also been used to deal with many complicated biosystems, e.g., the QSAR study (Prado-Prado et al., 2008), hard bionetwork systems (Diao et al., 2007; Gonzalez-Diaz et al., 2008), hepatitis B viral infection (Xiao et al., 2006), HBV virus gene missense mutations (Xiao et al., 2005b), visual analysis of SARS-CoV (Wang et al., 2005), representation of complicated biological sequences (Xiao et al., 2005a), and identification of protein attributes (Xiao and Chou, 2007). Graphical approaches are a hot topic in biological and medical science. With improvements in graphical analysis capabilities, we can obtain a more in-depth insight than ever. For instance, in a graph of homologous relationships, the aforementioned scale-free feature only indicates distribution of the connectivity of vertex. Even if the distributions are identical, a network may have a specific characteristic that distinguishes it from other networks. Our interest is in identifying some in-depth features specific for homologous relationships.

To obtain in-depth and detailed features, a reasonable standard for the definition of a homologous relationship is a prerequisite. Structure and sequence are two significant characteristics of a protein. Since remote homologous proteins can share the representative fold of a family, structure is a more robust characteristic than sequence. On the other hand, sequence similarity is vital in identifying homologous relationships. As we draw a network of homologous relationship, if only structural similarity is considered, proteins without similar biological properties might be mistakenly connected by an edge. This will result in a false detail in the graph. Similarly, when only sequence information is considered severe differences in biology properties (e.g., structure) might be tolerated. Thus, joint consideration of sequence and structural similarities is most appropriate for plotting a graph of homologous relationships (Qian et al., 2001).

Biological systems have evolved from simple to complex and from small to large. It has been proposed that short segments of polypeptides may have collapsed together to form folded proto-domains in the early evolution of proteins (Trifonov and Berezovsky, 2003; Riechmann and Winter, 2006). Domains evolved to their modern size through the assembly and/or exchange of smaller gene segments encoding polypeptide segments of sub-domain size (Blake, 1978), for example, by exon shuffling (Gibert, 1978) or non-homologous recombination (Bogarad and Deem, 1999). Thus, homologous relationships for short polypeptide segments represent an ideal aspect to investigate protein evolution. On the other hand, in protein evolution, insertion and deletion often occur in variable region, but to a lesser degree in conserved regions that are important for biological properties. Consequently, much progress has been achieved by matching homologous proteins with ungapped residue segments on a site-by-site basis (Smith et al., 1990; Henikoff and Henikoff, 1991, 1992). Since the alignment of ungapped residue segments retains most of the information significant for corresponding homologs, a suitable representation in characterizing homologous relationships for short polypeptide segments is also provided.

In the present study we propose a novel approach to investigate homologous relationships for proteins that provide useful information for various conformational diseases. We used information on ungapped aligned residue segments to plot a general graph of polypeptide relationships (GPR) by jointly considering sequence and structural similarities. Detailed analysis of the graph revealed a donut-shaped fingerprint in a vital subnetwork of the GPR. Using the information provided by this fingerprint, we identified switch sites for conformational conversion of prion, and other conformational disease-related proteins.

2. Methods

We investigated homologous relationships between pairs of residue segments. In total, 1612 non-membrane proteins from PDB_SELECT25 ([issued on 25 September 2001]Hobohm and Sander, 1994) were used in the analysis. In this non-redundant data set, no pair of sequences shares sequence identity of more than 25%. The solvent-accessible area for each residue was calculated using the DSSP (Kabsch and Sander, 1983) algorithm for every protein. A protein sequence is treated as a succession of residue segments. As the residue–residue correlation is notable in 15-residue segments (Liu et al., 2003), we used a window width of 15 for further consideration. By sliding a 15-residue window along the protein sequence, each segment of the data set serves as a query segment.

2.1. Search for remote homologs

In the universe of residue segments, samples are biased, i.e., some segments are closely related to others. To reduce this bias and to filter redundant samples and obtain a non-redundant GPR, we constructed a non-redundant target set of homolog searches $\{CR_{\leqslant 4}\}_{\mu}$ for each query segment μ. This target set is a subset of our database in which each segment shares no more than four common residues ($CR_{\leqslant 4}$, sequence identity is 26.7%) with the query segment. For each query, homologs are searched in the corresponding target set. In this way, all the segments obtained are remote homologs of the query polypeptide.

For each query segment μ, we searched the corresponding target set $\{CR_{\leqslant 4}\}_{\mu}$ for ungapped segments, i.e., remote homologs that are similar to the query in terms of both sequence and structure. This was carried out in two steps. First, multi-aligned remote homolog candidates of the query segment were initialized using a center-star approach. Then the remote homolog candidates were optimized using a position-specific matrix, an updated scoring scheme used in evaluating homologous relationships.

Definition of non-redundant structural analogs

For each query segment μ, if the following two conditions are satisfied, we say that segment v, $v \in \{CR_{\leqslant 4}\}_{\mu}$, is a non-redundant structural analog of μ.

1. Structural similarity $drms(\mu, v) < 4\,\text{Å}$, where the distance root mean squared deviation (*drms*; Park and Levitt, 1995)

for structure μ and v is defined as the average distance difference

$$drms(\mu,v)=\left[\frac{2}{15(15-1)}\sum_{i=2}^{15}\sum_{j=1}^{i-1}(|r_{\mu i}-r_{\mu j}|-|r_{vi}-r_{vj}|)^2\right]^{1/2}, \quad (1)$$

where r_{ai} is the coordinate of the C_α atom i in structure a.

2. Difference in surface residue $Z=\sum_{i=1}^{15}\delta(\varepsilon_{\mu i},\varepsilon_{vi})$ is at most 2 between μ and v, where $\varepsilon_{ai}=1$ for a surface residue and $\varepsilon_{ai}=0$ otherwise, $\delta(x,y)$ is a step function with $\delta(x,y)=0$ for $x=y$ and $\delta(x,y)=1$ otherwise. The contribution of a residue to the folding mechanism and protein function depends on whether or not it is exposed to solvent. To identify highly related samples, we investigated segments with similar exposed/buried residues to μ. A surface residue is defined as one with accessible area greater than 10% of the maximum accessible surface area (Chotia, 1975) for that type of residue.

Initialization step

Non-redundant structural analogs of query segment μ were searched in set $\{CR_{\leqslant 4}\}_\mu$. In addition to structural similarity, sequence similarity can be evaluated by the knowledge of the propensity of residues to substitute for each other in homologous proteins (Henikoff and Henikoff, 1992). Each non-redundant structural analog v of the query segment μ has a pairwise sequence alignment score $T(\mu,v)=\sum_{i=1}^{15}Score(\mu_i,v_i)$, where $Score(a_i,b_i)$ is an element of the BLOSUM30 matrix for the substitution of residues a_i and b_i. T is a measure of the degree of sequence similarity, and approximately corresponds to the biological homophyly between the non-redundant structural analog and the query segment. We ranked the non-redundant structural analogs of μ in descending order of score T, and used the top 30 samples as the initial remote homolog candidates in multiple sequence alignment of the subsequent optimization step.

Optimization step

In this step, a position-specific matrix of multiple sequence alignments was calculated for the top 30 samples. (The first matrix was calculated using sequences provided by step 1. Updated matrices were calculated using samples produced by step 2.) The scores for a specific position i are of the form

$$Profile(i,j)=\log\left(\frac{q_{ij}}{p_j}\right). \quad (2)$$

The probability of finding residue j in column i is estimated as $q_{ij}=(\alpha f_{ij}+\beta g_j)/(\alpha+\beta)$, where f_{ij} is the observed frequency for residue j, α and β are the relative weights for the observed and pseudocount residue frequencies, $\alpha=N$ is the number of aligned segments, and β is reasonably set as $\beta=10$. The pseudocount frequencies $g_j=\sum_k(f_{ik}/p_k)\tilde{q}_{jk}$, $\tilde{q}_{jk}$ are the target frequencies according to data from the BLOSUM30 matrix. p_l is the background probability of the occurrence of residue l implicit in the BLOSUM30 matrix. For a segment τ, the alignment score is calculated as

$$T'(\tau)=\sum_{i=1}^{15}Profile(i,\tau_i). \quad (3)$$

We searched target set $\{CR_{\leqslant 4}\}_\mu$ for non-redundant structural analogs (structural similarity retained) of the query segment μ and ranked them in decreasing order of score T' (sequence similarity retained). The top 50 segments were identified as updated remote homolog candidates. After reranking these 50 segments using the HFnet (Hydrophobic Force network) algorithm, which boosts the quality of sequence alignment (see supporting information), the position-specific matrix was updated with the top 30 samples. The contribution from HFnet was nearly convergent after seven iterations. Thus, to maximize the alignment quality, a total of seven iterations were processed. Then the final top 30 segments were selected as remote homologs, and formed the remote homolog set $\{R_\mu\}$ for polypeptide μ.

In general, three factors are considered in scanning remote homologs:

1. Sequence identity is limited to 26.7%, so that a non-redundant graph is obtained.
2. Structural similarity is required during the initialization and optimization processes, so that structural information is not lost.
3. Sequence similarity is retained by adopting the top-ranked samples according to the BLOSUM30 homolog database and the updated position-specific matrix.

2.2. GPR construction

We attempted to plot a homologous relationship network for the whole universe of polypeptides. Each query segment of our database was defined as a node of the polypeptide relationship network. According to the above method, remote homologs were found for each of these nodes/queries. Two nodes A and B are deemed to be related if $B\in\{R_A\}$ or $A\in\{R_B\}$, where $\{R_A\}$ and $\{R_B\}$ are the remote homolog sets for A and B, respectively. If $\{R_A\}$ and $\{R_B\}$ share no less than five segments, we say that nodes A and B are connected by edge (A,B). Owing to our definition, each pair of connected nodes/polypeptides has similar biological properties, but a low level of sequence identity. Consequently, we constructed a non-redundant unweighted GPR in which each edge is considered equally. For each node, the value of connectivity K is defined as the number of edges connected to the node.

To decrease the probability of false connection, we introduced an optimization approach by counting the shared segments between two remote homolog sets. The homologous relationship is credible if A shares enough remote homologs with B. A low threshold for the number of shared segments results in a high level of false connections, whereas a high threshold results in more orphans. Empirically, we recommend 5 as a threshold.

3. Results

3.1. Donut fingerprint

As shown in Fig. 1, the polypeptide population fits a power law. Other than this approximate feature, further knowledge has seldom been mentioned in the literature. In fact, the network of homologous relationships in the polypeptide universe is so vast and complicated that many researchers have avoided a detailed analysis. Consequently, to the best of our knowledge, detailed features of the network of polypeptide relationships are still unknown. Here we investigated such details by analyzing the GPR character from a vital subgraph formed by significant vertices. Using PAJEK software (http://vlado.fmf.uni-lj.si/pub/networks/pajek/,), the representative features of a network can be analyzed by topological transformation. In this algorithm, nodes and edges are placed in a plane. Relative nodes are close to each other by introducing a virtual attracting force between vertices connected

by an edge. On introduction of a virtual repulsive force, all vertices are repelled from each other so that no pair of vertices can get too close. The topological structure of a network is transformed by minimizing the system energy. The coordinates of node clusters converge after energy minimization in PAJEK. As a result of such topological transformations, tightly correlated nodes, i.e., homologous polypeptides, are bunched into node clusters.

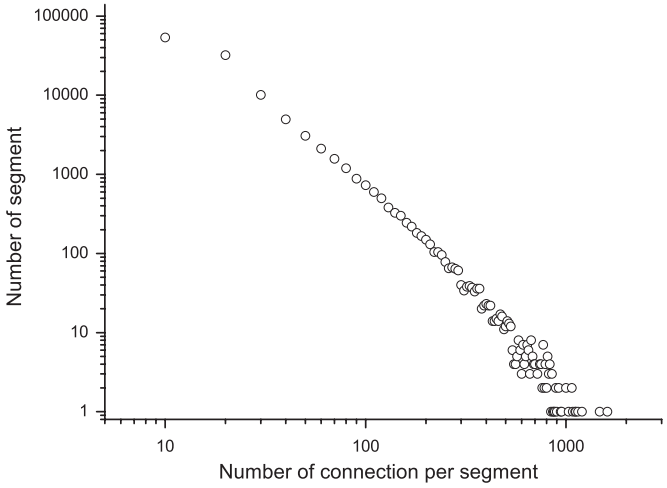

Fig. 1. Node population as a function of connectivity.

We plotted subnetworks of the polypeptide relationship for nodes with connectivity $K>60$. The vertices of the network were colored according to the protein secondary structure (taken from DSSP database. As in most methods, we considered three types of conformation $\{h,e,c\}$ generated from the eight possible by coarse graining of $h,g,i\rightarrow h$, $e\rightarrow e$ and $x,t,s,b\rightarrow c$). In a polypeptide, a subsegment $a_ia_{i+1}a_{i+2}a_{i+3}a_{i+4}a_{i+5}a_{i+6}$ ($i=0$ or 8) is defined to be **H** if more than three of its residues are in helix conformation, **E** if more than three of its residues are in strand conformation, and **C** otherwise. Thus, nine non-overlapping polypeptide states are defined: **HH**, **HC**, **CH**, **EE**, **EC**, **CE**, **HE**, **EH**, and **CC**. The graphs obtained for these subnetworks are shown in Fig. 2 in decreasing order of connectivity. A clear donut-shaped fingerprint is evident. The Pajek software includes several options for clustering that differ in force model and distance measure. We tested many different options. The resulting donut-shaped topological feature is robust.

Helix segments and N- and C-terminal caps (**HH** + **HC** + **CH**) make up the main body of the donut shape. Significant groups/types of strand segments (and their caps) are not connected to the ring or to each other in subnetworks $K>100$. When nodes with connectivity of up to 80 are considered (Fig. 2E,F), such connections emerge with decreasing K. As shown in Fig. 2E, nodes that connect the diameter of the donut shape appear at approximately $K=60$. With a further decrease in K, crossings between different parts of the donut ring appear. In other words, the ring in the GPR is connected by nodes with low connectivity.

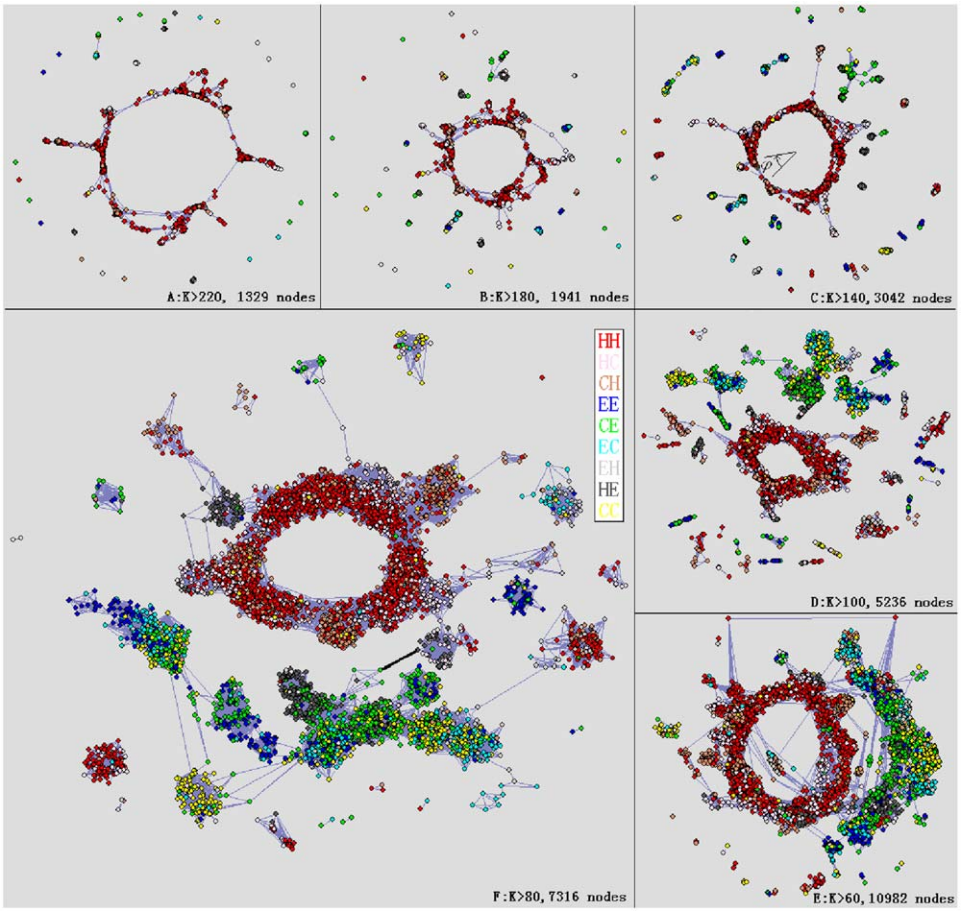

Fig. 2. Donut-shaped fingerprint of the polypeptide relationship network. Nodes with a connection number K greater than 60 are plotted. Orphans in these subnetworks are omitted. Tightly related nodes are bunched up by PAJEK. The donut is rich in **HH** + **HC** + **CH** samples. The arc in E is rich in **EE** + **EC** + **CE** samples. A connected strand-arc is evident in F. In F, there is only one edge (colored in black) that connects the helix-donut to strand-arc.

Table 1
Residue coverage of the data set given by a subnetwork of the graph of polypeptide relationships.

Connectivity threshold	Count-NON	Coverage-NON (%)	Coverage-SRNON (%)	Coverage-SCNON (%)
K>220	1329	1.7	85.2	74.1
K>180	1941	2.5	91.2	80.9
K>140	3042	3.7	96.0	87.7
K>100	5236	5.9	97.9	91.7
K>80	7316	7.9	98.5	93.5
K>60	10982	11.1	99.0	95.2

Count_NON, number of no-orphan nodes in a subnetwork; Coverage-NON, coverage given by no-orphan nodes in a subnetwork; Coverage-SRNON, coverage given by self and related nodes of the no-orphan nodes in a subnetwork (if $B \in \{R_A\}$ or $A \in \{R_B\}$, A and B are related); Coverage-SCNON, coverage given by self and directly connected nodes of the no-orphan nodes in a subnetwork.

We find that the subgraphs shown in Fig. 2 are not trivial node-limited profiles of polypeptide relationships, but characterize the topological feature of the whole GPR. As shown in Table 1, for nodes with connectivity of $K>80$, 7316 no-orphan nodes ($K\neq 0$) exist in the corresponding subnetwork (Fig. 2F). Although these nodes/segments are contributed by only 7.9% of the amino acids in our database, related (according to the definition in Subsection 2.2) or directly connected (two nodes directly connected by one edge) segments of these nodes cover residues of nearly the whole data set. As these segments have similar biological properties to the corresponding nodes, it means that the aforementioned simple topological feature represents the nature of the whole polypeptide universe, i.e., there are two nearly separated regions in phase space of polypeptide segment: a helix-donut zone and a strand-arc zone. The two parts are connected flimsily by sparse edge. Although we cannot draw a picture of the whole GPR because of its extreme complexity, and only vital subgraphs can be depicted, the position of a segment in phase space of polypeptide can be deduced from secondary structure. We assumed that **HH** + **HC** + **CH** samples belong to the helix-donut zone, whereas **EE** + **EC** + **CE** samples are in the strand-arc zone. Then a picture of the whole graph of the polypeptide universe is constructed. Moreover, the origin of the complicated protein universe might be very neat. As shown by the first two rows of Table 1, nodes shown in Fig. 2A,B, comprising approximately 2% of the residues in our database, 'determine' the properties of nearly 80–90% of the sites in the database.

To reveal the reason for the donut shape, we selected the shape shown in Fig. 2C, a network of moderate complexity, for detailed analysis. In this graph, the coordinates of node clusters represent the approximate position of a specific group of homologous polypeptides. By introducing a virtual center and a clockwise angle φ, samples of the donut shape in successive $\pi/6$ slices were investigated. Polypeptides in each slice were matched site by site. The probability densities for buried and hydrophobic residue were calculated for each site (residue classification was: hydrophobic h = {M, F, I, L, V, A, W}, polar p = {C, Y, Q, H, P, G, T, S, N, R, K, D, E} (Liu et al., 2002)). As shown in Fig. 3, with the variation of φ, a successive shift in buried/hydrophobic residues was observed for polypeptides making up the donut shape. Thus, the distribution of buried/hydrophobic residues may be closely related to the donut shape. Since helix forms are abundant in this shape, the period of buried/hydrophobic residues is approximately 4.

3.2. Switch sites for prions

Moderate conversion of their structure is vital for the biological properties of protein molecules. Thus, moderate

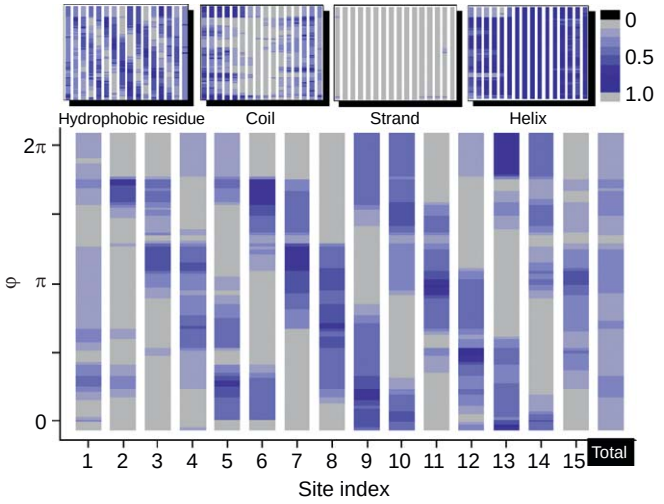

Fig. 3. Buried residues of a donut ring for $K>140$ as a function of φ. Samples in each pie slice are matched site by site. At each site, the probability of a buried residue is shown by different brightness. The insert shows the distribution of hydrophobic residues and secondary structures obtained by the same method.

changes in the structure of homologous protein are allowable, whereas a significant conversion may not be possible. As illustrated by vital subgraphs (Fig. 2E,F), there are a limited number of nodes in the whole GPR that form a 'bridge' between the helix-donut zone and the strand-arc zone. This means that, in terms of protein evolution, significant structural conversion, e.g., a change from a helix to a sheet, is difficult. Consequently, the protein universe is in a relatively steady state, with infrequent exceptions. One well-known exception is the prion protein (PrP) that exhibits a change in structure in pathological conditions (Prusiner, 1982, 1998) .

PrP is deemed to be responsible for transmissible spongiform encephalopathies (TSEs), a group of fatal neurodegenerative diseases that are associated with conformational conversion of the normally monomeric and α-helical protein molecule, PrP^{C}, to the β-sheet-rich PrP^{Sc}. TSEs arise in several mammalian species by genetic, infectious, or sporadic means, and include bovine spongiform encephalopathy in cattle, scrapie in sheep, chronic wasting disease in cervids, and Creutzfeldt–Jakob disease and kuru in humans (Prusiner, 1982, 1998; Caughey and Baron, 2006; Collinge, 2001; Aguzzi and Polymenidou, 2004; Weissmann, 2004). It is now widely accepted that in these protein-only diseases (Prusiner, 1982), TSE transmission does not require nucleic acids.

As a marginally stable form between α-helix-rich and β-sheet-rich states, PrP must be a protein with inbuilt polypeptides related to some 'bridge' nodes of the GPR (nodes connecting the helix-donut zone to the strand-arc one). It is also reasonable that such inbuilt polypeptides should correlate with the origin of conformational conversion. Although identifying a detailed mechanism for this structural conversion is beyond the scope of the present study, and the structure of PrP^{Sc} is also largely unknown, we can apply our view of protein evolution to identify the segment in which the conformational conversion arises.

By sliding a window along the residue sequence, each 15-residue segment of human PrP (hPrP, 121–230, PDB ID:1QM2) was analyzed in terms of GPR. Vertices and edges in whole GPR were used as a framework for defining polypeptide relationships. For a given segment of hPrP, the top 30 remote homologs were searched in whole GPR with the method described in Subsection 2.1. By definition, these remote homologs are highly similar to the query hPrP segment in sequence and structure, i.e., they represent

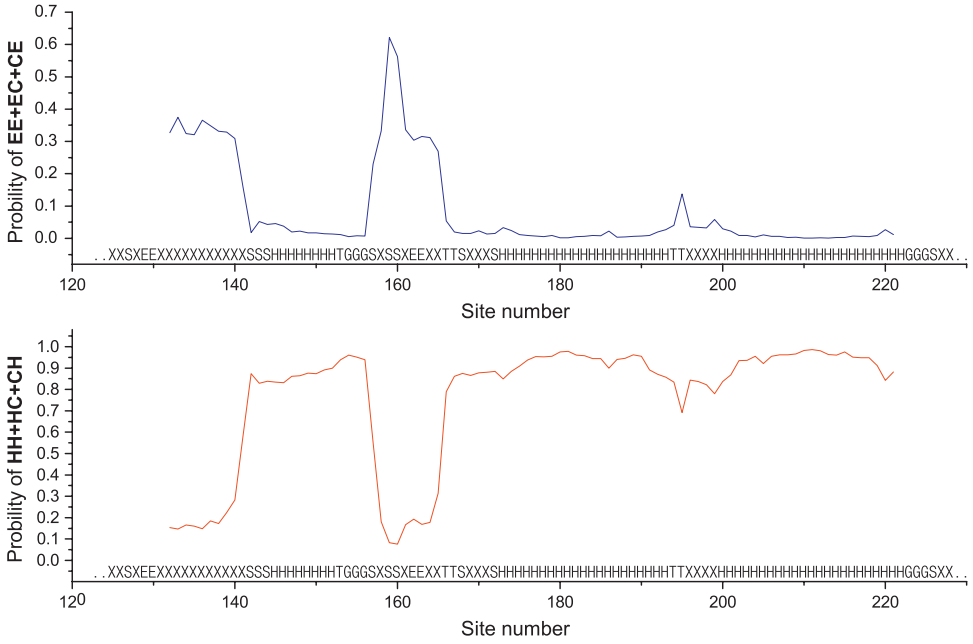

Fig. 4. Frequency of two types of nodes connected to agents of the query segments of human prion protein 1QM2.

agents of the query segment. Nodes of the GPR directly connected to these agents were identified. States (**HH**, **HC**, etc.) of these collected nodes indicate the probability of whether an agent belongs to a helix-donut zone (corresponding to **HH** + **HC** + **CH**) or a strand-arc zone (**EE** + **EC** + **CE**). We assigned identified nodes to sites of the central residue of the query hPrP segment. The frequencies of the types of nodes identified are shown in Fig. 4 site by site.

Usually the aggregation-prone regions tend to be blocked in native state of globular proteins because side chains are hidden in the inner hydrophobic core, or the cellular environment forbids the condition of the formation of aggregation (Dobson, 1999). A nosogenetic misfolding starts in a region where the unlocking begins. It is like a switch of exposure of sensitive regions. Then based on an exposure or partly exposure (Claudio, 2001), the aggregation-prone regions might have the further chance to form amyloid in the following folding pathway. Here we attempted to predict such switch sites in hPrP using the GPR feature of sparse connections between the helix-donut zone and the strand-arc. We assumed that conformational conversion is due to transition between two regions of polypeptide phase space. If polypeptides change their structures near native conformations, there will be no conformational disease. So we should pay attention to segments which are prone to fold to structures other than their native conformations. In Fig. 4, except for sites of two inborn strands of approximately 130 and 160, there is a peak for **EE** + **EC** + **CE** at site 195. With a high frequency for **EE** + **EC** + **CE** and a low frequency for **HH** + **HC** + **CH**, the two inborn strands can easily extend according to the GPR. Due to thermal motion, a protein molecule can moderately change its conformation at room temperature. Such facile extension of the inborn strands should be allowed by PrP^C, otherwise, if it could cause disease, the corresponding life-form would have been lost during evolution. Therefore, it is likely that such a site is not responsible for conformational changes related to disease. On the other hand, sites around position 195 are different. As shown in Fig. 4, the native conformation of this region is in **HH** + **HC** + **CH**. As these sites have a high probability of being in their inborn helix-donut state, normally it is difficult to change state to a strand-arc. While in this special case there is reasonable probability that the polypeptide will transform to the strand-arc region, i.e., induce a conformational conversion. Consequently, residues around position 195 (≈ 188–202) should be responsible for the disease-related conformational change. This conclusion contrasts to earlier, largely theoretical models, and is consistent with the experimental observations of Kuwata et al. (2007) that intercalation of an anti-prion compound GN8 to regions N159, V189, T192, K194, and E196 hampers the pathogenic conversion process. Moreover, as non-redundant polypeptides were used throughout our approach and analysis, this conclusion should be the same for all members of the PrP family.

4. Discussion

Here we identified a simple feature of the evolution of protein molecules, and presented a general picture of the non-membrane polypeptide universe. In the GPR there are few shortcuts connecting the diameter of a donut and 'bridges' between the helix-donut zone and the strand-arc. Such crossing nodes are of low connectivity. This indicates that homologous relationships generally evolve gradually. Most polypeptides evolved strictly along a helix-donut or a strand-arc track, with very few samples exhibiting a drastic shift in biological properties during evolution, e.g. as shown in Fig. 3, such an evolvement induces a gradual change in the distributions of buried and hydrophobic residue and thus in biochemical properties. While it is interesting that the evolvement can final hook-up and form a ring. Since the present work focuses on divergent evolution, it suggests that divergent evolution can result in convergent evolution at a sub-domain level, but in a gradual way that induces a donut-shaped topological feature.

It is interesting to make a second consideration of the formation of donut ring. Shift of the distributions of buried and hydrophobic residue has a high correlation with donut. While with the evolvement of polypeptide segment, there should be opportunity to form different groups of polypeptide structures. Each group owns a donut-shaped fingerprint with shift of

buried/hydrophobic residue, but a different way in the change of three-dimensional structure. This would result in several connected rings. But it does not happen. As the GPR is a network with moderate structural deviation, a two-step evolvement may arouse severe structural change as big as that among different groups. This can be illustrated by the insertions in Fig. 3, where there is no obvious character in the distribution of protein secondary structure. Consequently, the candidate donut rings final joint together, and only one ring is resulted. In such a consideration, graph of the enlarged polypeptide segment will only correspond to further shift of the distribution of buried/hydrophobic residue, and the one ring donut-shaped fingerprint will reoccur. Actually, we have drawn the graph of 17 and 19 length segments, and have found similar fingerprints too.

A marked difference between this study and others is that the picture obtained not only provides details of topology features, but also has direct and important applications. According to GPR, sparse connection between the helix-donut and the strand-arc is an indicator of conformational changes related to disease. As shown by the analysis of PrP, we can use conformational information on one state to deduce the switch sites for structural conversion related to pathological conditions. This study can be extended to other conformational diseases, such as sickle cell anemia, antithrombin deficiency thromboembolic disease, and familial amyloid neuropathy (see supporting information; details to be published elsewhere). Identification of the site of origin of such conformational conversions is extremely important in designing suitable therapy approaches. By revealing switch sites for structural conversion, we can design drugs to hamper this pathogenic conversion process, or even upgrade species by mutation. Such switch sites are usually determined by cases in which the switch role is evident, such as disease-related point mutations reported in clinic and experiments in hampering the pathogenic conversion process. The cost of such research is considerable. More significantly, the disease conformation was believed to be a prerequisite in previous research, which limited the number of proteins that could be investigated. A systemic study of conformational diseases in organisms was thus beyond the scope of previous approaches. The knowledge provided by GPR can be used to overcome the requirement for unnecessary disease structures, to predict target sites for clinical treatment, and to investigate suitable therapy schemes based on normal proteins. This new approach could lead to great progress in curing conformational diseases. Moreover, as demonstrated by the example described here, GPR considers both structural information and sequence identity, and thus represents a suitable strategy for meeting challenges in the design of conformational protein switches (Ambroggio and Kuhlman, 2006).

Connections in GPR are highly exact. As shown in Fig. 2F, none of the 109,045 edges of the subgraph make a false connection crossing the diameter. As our aim was to provide a general picture of the whole universe of polypeptide segments, nodes and connections should be both representative and properly weighted so that the resulting feature is universal but not biased. Consequently remote homologous relationship was selected as a feasible representation. While in this representation, if the criterion for structural similarity is too strict, there will be a drastic decrease in the number of suitable candidates. Thus we set the cut-off as $drms < 4$ Å, which is a moderate level. This provides the opportunity for false connection. However, such false connections are finally controlled. This owed much to the HFnet algorithm. In 2008 we suggested that the family representative intramolecular hydrophobic force networks makes a crucial contribution to the biological properties conserved throughout protein evolvement (Liu et al., 2008a). It uncovers the truth of protein evolution significantly. Based on this theory, we have developed a model called HFnet to evaluate the significance of each sequence in a given multiple sequence alignment. The power of HFnet has been proven not only in silico, but also in wet experiment. Based on the HFnet algorithm, we have ever designed five artificial remote proteins of the WW domain. As all of them have low pairwise sequence identity ($<30\%$) with each other and with each proteins in the learning set, it is usually difficult to write out such sequences, and say nothing of a family sharing specific biological properties. However, in biological experiment, four of them exhibited detectable ligand-binding affinity. These experiment data demonstrated that our theory and the HFnet algorithm are very robust, and dominate/identify not only protein structure but also biological properties. In the present case, HFnet algorithm contributed at least 50% increase in accuracy of remote homologs identification. However, as only two letters were used in HFnet, for such a simple algorithm, signals for segments that are too short may be missed. This was another consideration when selecting the 15-residue window. Fortunately, as structural information was also considered in this work, a 15-residue polypeptide was long enough for HFnet. With a decrease in residue–residue correlation (Liu et al., 2003), a greater window length would cover more secondary factors. However, as there are only two major conformations, a helix and a strand, in protein molecule, the donut–arc topological feature should not be remarkably modified.

As we have minimized the false signals during network construction, the vertices and connections in the GPR can be used as a framework that reliably represents the universe of polypeptide relationships. The biological properties of a protein can be credibly predicted from such a representation. This will facilitate studies of complex proteins and allow noise-free analysis. Further improvements and applications of this representation are currently being investigated.

Acknowledgments

We are grateful to professors Zeng-Ru Di, Zuo-Bing Wu, and Yuan-Kai Hong, and doctors Fang-Ting Li and Ming Li for their helpful discussions. To ensure a healthy development of modern biology, a patent is applied for corresponding method. We encourage pure scientific research. Contact authors when the method is to be used. This work was jointly supported by the National High-tech R&D Program of China (863 Program, Grant No. 2007AA021803), National Basic Research Program of China (973 Program, Grant No. 2007CB310500), and National Natural Science Foundation of China, No. 10704077.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at 10.1016/j.jtbi.2009.02.009.

References

Aguzzi, A., Polymenidou, M., 2004. Mammalian prion biology: one century of evolving concepts. Cell 116, 313–327.

Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1993. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. J. Biol. Chem. 268, 6119–6124.

Ambroggio, X.I., Kuhlman, B., 2006. Design of protein conformational switches. Curr. Opin. Struct. Biol. 16, 525–530.

Andraos, J., 2008. Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. Can. J. Chem. 86, 342–357.

Blake, C.C.F., 1978. Do genes-in-pieces imply proteins-in-pieces? Nature 273, 267.

Bogarad, L.D., Deem, M.W., 1999. A hierarchical approach to protein molecular evolution. Proc. Natl Acad. Sci. USA 96, 2591–2595.

Caughey, B., Baron, G.S., 2006. Prions and their partners in crime. Nature 443, 803–810.
Chotia, C., 1975. The nature of the accessible and buried surfaces in proteins. J. Mol. Biol. 105, 1–14.
Chou, K.C., 1989. Graphical rules in steady and non-steady enzyme kinetics. J. Biol. Chem. 264, 12074–12079.
Chou, K.C., 1990. Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophys. Chem. 35, 1–24.
Chou, K.C., 1992. Energy-optimized structure of antifreeze protein and its binding mechanism. J. Mol. Biol. 223, 509–517.
Chou, K.C., 2004a. Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. Biochem. Biophys. Res. Commun. 316, 636–642.
Chou, K.C., 2004b. Structural bioinformatics and its impact to biomedical science. Cur. Med. Chem. 11, 2105–2134.
Chou, K.C., Forsen, S., 1980. Graphical rules for enzyme-catalyzed rate laws. Biochem. J. 187, 829–835.
Chou, K.C., Zhang, C.T., 1992. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Res. Hum. Retrov. 8, 1967–1976.
Chou, K.C., Shen, H.B., 2007a. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem. Biophys. Res. Commun. 357, 633–640.
Chou, K.C., Shen, H.B., 2007b. Recent progresses in protein subcellular location prediction. Anal. Biochem. 370, 1–16.
Chou, K.C., Shen, H.B., 2008. Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat. Protoc. 3, 153–162.
Chou, K.C., Kezdy, F.J., Reusser, F., 1994. Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. Anal. Biochem. 221, 217–230.
Claudio, S., 2001. Protein misfolding and disease; protein refolding and therapy. FEBS Lett. 498, 204–207.
Collinge, J., 2001. Prion disease of human and animals, their cause and molecular basis. Annu. Rev. Neurosci. 24, 519–550.
Diao, Y., Li, M., Feng, Z., Yin, J., Pan, Y., 2007. The community structure of human cellular signaling network. J. Theor. Biol. 247, 608–615.
Dobson, C.M., 1999. Protein misfolding, evolution and disease. Trends Biochem. Sci. 24, 329–332.
Dokholyan, N.V., Shakhnovich, B., Shakhnovich, E.I., 2002. Expanding protein universe and its origin from the biological Big Bang. Proc. Natl Acad. Sci. USA 99, 14132–14136.
England, J.L., Shakhnovich, B.E., Shakhnovich, E.I., 2003. Natural selection of more designable folds: a mechanism for thermophilic adaptation. Proc. Natl Acad. Sci. USA 100, 8727–8731.
Finkelstein, A.V., Gutin, A., Badretdinov, A., 1995. Why do protein architectures have Boltzmann-like statistics? Proteins 23, 142–150.
Gibert, W., 1978. Why genes in pieces? Nature 271, 501.
Gonzalez-Diaz, H., Gonzalez-Diaz, Y., Santana, L., Ubeira, F.M., Uriarte, E., 2008. Proteomics, networks, and connectivity indices. Proteomics 8, 750–778.
Govindarajan, S., Goldstein, R.A., 1996. Why are some protein structures so common? Proc. Natl Acad. Sci. USA 93, 3341–3345.
Henikoff, S., Henikoff, J.G., 1991. Automated assembly of protein blocks for database searching. Nucleic Acids Res. 19, 6565–6572.
Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl Acad. Sci. USA 89, 10915–10919.
Hobohm, U., Sander, C., 1994. Enlarged representative set of protein structures. Protein Sci. 3, 522–524.
Holm, L., Sander, C., 1993. Protein-structure comparison by alignment of distance matrices. J. Mol. Biol. 233, 123–138.
Holm, L., Sander, C., 1996. Mapping the protein universe. Science 273, 595–602.
Holm, L., Sander, C., 1997. An evolutionary treasure, unification of a broad set of amidohydrolases related to urease. Proteins 28, 72–82.
Huynen, M.A., van Nimwegen, E., 1998. The frequency distribution of gene family sizes in complete genomes. Mol. Biol. Evol. 15, 583–589.
Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure. Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.
Koonin, E.V., Wolf, Y.I., Karev, G.P., 2002. The structure of the protein universe and genome evolution. Nature 420, 218–223.
Kuwata, K., Nishida, N., Matsumoto, T., 2007. Hot spots in prion protein for pathogenic conversion. Proc. Natl Acad. Sci. USA 104, 11921–11926 (14 co-authors).
Kuzmic, P., Ng, K.Y., Heath, T.D., 1992. Mixtures of tight-binding enzyme inhibitors. Kinetic analysis by a recursive rate equation. Anal. Biochem. 200, 68–73.
Li, H., Helling, R., Tang, C., Wingreen, N.S., 1996. Emergence of preferred structures in a simple model of protein folding. Science 273, 666–669.
Li, H., Tang, C., Wingreen, N.S., 1998. Are protein folds atypical? Proc. Natl Acad. Sci. USA 95, 4987–4990.
Liu, X., Liu, D., Qi, J., Zheng, W.M., 2002. Simplified amino acid alphabets based on deviation of conditional probability from random background. Phys. Rev. E 66, 021906.
Liu, X., Zhang, L.M., Guan, S., Zheng, W.M., 2003. Distances and classification of amino acids for different protein secondary structures. Phys. Rev. E 67, 051927.
Liu, X., Zhang, L.M., Yin, J., Zhao, Y.P., 2008a. Major factors of protein evolution revealed by eigenvalue decomposition analysis. In: Proceeding of the International Conference on Bioinformatics and Computational Biology BIOCOMP'08, Las Vegas, USA, pp. 91–97.
Liu, X., Zhao, Y.P., Zheng, W.M., 2008b. CLEMAPS: multiple alignment of protein structures based on conformational letters. Proteins 71, 728–736.
Myers, D., Palmer, G., 1985. Microcomputer tools for steady-state enzyme kinetics. Comput. Appl. Biosci. 1, 105–110.
Orengo, C.A., Todd, A., Thornton, J.M., 1999. From protein structure to function. Curr. Opin. Struct. Biol. 9, 374–382.
Park, B.H., Levitt, M., 1995. The complexity and accuracy of discrete state models of protein structure. J. Mol. Biol. 249, 493–507.
Prado-Prado, F.J., Gonzalez-Diaz, H., de la Vega, O.M., Ubeira, F.M., Chou, K.C., 2008. Unified QSAR approach to antimicrobials. Part 3: First multi-tasking QSAR model for input-coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. Bioorgan. Med. Chem. 16, 5871–5880.
Prusiner, S.B., 1982. Novel proteinaceous infectious particles cause scrapie. Science 216, 136–144.
Prusiner, S.B., 1998. Prions. Proc. Natl Acad. Sci. USA 95, 13363–13383.
Qi, X.Q., Wen, J., Qi, Z.H., 2007. New 3D graphical representation of DNA sequence based on dual nucleotides. J. Theor. Biol. 249, 681–690.
Qian, J., Luscombe, N.M., Gerstein, M., 2001. Protein family and fold occurrence in genomes, power-law behavior and evolutionary model. J. Mol. Biol. 313, 673–681.
Riechmann, L., Winter, G., 2006. Early protein evolution: building domains from ligand-binding polypeptide segments. J. Mol. Biol. 363, 460–468.
Rost, B., 1997. Protein structures sustain evolutionary drift. Fold. Des. 2, s19–s24.
Russ, W.P., Lowery, D.M., Mishra, P., Yaffe, M.B., Ranganathan, R., 2005. Natural-like function in artificial WW domains. Nature 437, 579–583.
Shakhnovich, B.E., Deeds, E., Delisi, C., Shakhnovich, E., 2005. Protein structure and evolutionary history determine sequence space topology. Genome Res. 15, 385–392.
Shen, H.B., Chou, K.C., 2007. Signal-3L: a 3-layer approach for predicting signal peptide. Biochem. Biophys. Res. Commun. 363, 297–303.
Sirois, S., Wei, D.Q., Du, Q.S., Chou, K.C., 2004. Virtual screening for SARS-CoV protease based on KZ7088 pharmacophore points. J. Chem. Inf. Comput. Sci. 44, 1111–1122.
Smith, H.O., Annau, T.M., Chandrasegaran, S., 1990. Finding sequence motifs in groups of functionally related proteins. Proc. Natl Acad. Sci. USA 87, 826–830.
Socolich, M., Lockless, S.W., Russ, W.P., Lee, H., Gardner, K.H., Ranganathan, R., 2005. Evolutionary information for specifying a protein fold. Nature 437, 512–518.
Taverna, D.M., Goldstein, R.A., 2000. The distribution of structures in evolving protein populations. Biopolymers 53, 1–8.
Teichmann, S.A., Chothia, C., Gerstein, M., 1999. Advances in structural genomics. Curr. Opin. Struct. Biol. 9, 390–399.
Trifonov, E.N., Berezovsky, I.N., 2003. Evolutionary aspects of protein structure and folding. Curr. Opin. Struc. Biol. 13, 110–114.
Wang, M., Yao, J.S., Huang, Z.D., Xu, Z.J., Liu, G.P., Zhao, H.Y., Wang, X.Y., Yang, J., Zhu, Y.S., Chou, K.C., 2005. A new nucleotide-composition based fingerprint of SARS-CoV with visualization analysis. Med. Chem. 1, 39–47.
Weissmann, C., 2004. The state of the prion. Nat. Rev. Microbiol. 2, 861–871.
Wolynes, P.G., 1996. Symmetry and the energy landscapes of biomolecules. Proc. Natl Acad. Sci. USA 93, 14249–14255.
Wong, P., Frishman, D., 2006. Fold designability, distribution, and disease. PLoS Comp. Biol. 2, 0392–0402.
Xiao, X., Chou, K.C., 2007. Digital coding of amino acids based on hydrophobic index. Protein Pept. Lett. 14, 871–875.
Xiao, X., Shao, S.H., Chou, K.C., 2006. A probability cellular automaton model for hepatitis B viral infections. Biochem. Biophys. Res. Commun. 342, 605–610.
Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., Chou, K.C., 2005a. Using cellular automata to generate image representation for biological sequences. Amino Acids 28, 29–35.
Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., Chou, K.C., 2005b. An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. J. Theor. Biol. 235, 555–565.
Yanai, I., Camacho, C.J., DeLisi, C., 2000. Predictions of gene family distributions in microbial genomes, evolution by gene duplication and modification. Phys. Rev. Lett. 85, 2641–2644.
Zhang, C.T., Chou, K.C., 1994. Analysis of codon usage in 1562 *E. coli* protein coding sequences. J. Mol. Biol. 238, 1–8.
Zhang, L.M., Liu, X., 2008. Significant residue features revealed by eigenvalue decomposition analysis of BLOSUM matrices. Phys. Lett. A 372, 2282–2285.
Zhou, G.P., Deng, M.H., 1984. An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. Biochem. J. 222, 169–176.
⟨http://vlado.fmf.uni-lj.si/pub/networks/pajek/⟩.