Vol 16 No 2, February 2007
1009-1963/2007/16(02)/0392-13

Chinese Physics

© 2007 Chin. Phys. Soc.
and IOP Publishing Ltd

Protein structural codes and nucleation sites for protein folding

Jiang Fan(江　凡)† and Li Nan(李　南)

Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing 100080, *China*

(Received 5 June 2006; revised manuscript received 31 August 2006)

One of the long-standing controversial arguments in protein folding is Levinthal's paradox. We have recently proposed a new nucleation hypothesis and shown that the nucleation residues are the most conserved sequences in protein. To avoid the complicated effect of tertiary interactions, we limit our search for structural codes to the nucleation residues. Starting with the hypotheses of secondary structure nucleation and conservation of residues important for folding, we have analysed 762 folds classified as unique by SCOP. Segments of 17 residues around the top 20% conserved amino acids are analysed, resulting in approximately 100 clusters each for the main secondary structure classes of helix, sheet and coil. Helical clusters have the longest correlation range, coils the shortest (four residues). Strong specific sequence-structure correlation is observed for coil but not for helix and sheet, suggesting a mapping relationship between the sequence and the structure for coil. We propose that the central sequences in these clusters form 'structural codes', a useful basis set for identifying nucleation sites, protein fragments stable in isolation, and secondary structural patterns in proteins (particularly turns and loops).

Keywords: nucleation, hydrogen bond, secondary structure, structural code, protein folding, sequence-structure relationship

PACC: 3620, 8715B, 8715D

1. Introduction

The sequence-structure relationship is an important aspect of protein folding studies. The sequence-structure relationship refers to a quasi-deterministic mapping between the amino acid sequence of a protein and its three-dimensional (3D) atomic structure. Although there is a probabilistic nature to this relationship, the principle, first established by Anfinsen,[1] that the 3D structure of aprotein is determined by its amino acid sequence, still holds for most proteins. The reason that this relationship holds is that the native structure has a free energy close to its global minimum. This is called the thermodynamic principle of protein folding. In company with the thermodynamic principle, there is the kinetic principle. The difficulty encountered in deciphering the kinetic principle has been pointed out in the Levinthal paradox.[2] The key problems are as follows. How is an unfolded protein in an extended conformation folded into a compact globular protein? What role does each amino acid play in specifying the folding pathway and the final folded structure? Does a second-half of the genetic code, namely, the structural code in proteins exist? If we could formulate a set of rules that are specific, universally applicable to the sequence-structure relationship in proteins, it will greatly deepen our understanding of protein folding and facilitate protein structure prediction and functional design.

Early attempts to look for the sequence-structure code primarily focused on the secondary structures of proteins.[3] What complicates the problem is that some of the secondary structures are determined by short-range interactions and others by long-range interactions. The sequence-structure patterns affected by the long-range interactions are in general involved with many-body interactions and thus are hard to delineate. We recently proposed a new folding theory based on a secondary structure nucleation hypothesis.[4,5] The nucleation hypothesis states that folding starts with a set of nucleation residues whose secondary structures are dominated by the short-range interactions, namely, by the flanking sequences. According to this hypothesis, the nucleation residues should be surrounded by the most conserved amino acid sequences in proteins, and furthermore, similarly conserved sequences should exist among all globular proteins. Based on this hypothesis we have designed

†E-mail: fjiang@aphy.iphy.ac.cn

a new nearest-neighbour method of predicting protein secondary structures and shown that we can achieve a higher prediction accuracy for residues with more conserved sequences.[4] More interestingly, the nucleation residues, which are defined by selecting the top 20% most conserved sequences, display a surprising pattern of hydrophobic residues more buried and hydrophilic residues more exposed. In contrast, this pattern cannot be observed by examining all (100%) residues in proteins, which means that the top 20% most conserved residues play the role of the core residues responsible for both driving and stabilizing the protein folding process.

Furthermore, for supporting our nucleation hypothesis, we have observed two scaling laws in proteins and derived the critical exponents for the two scaling laws by modelling them with two continuous phase transitions of finite size.[5] It is known that polymer statistics can be described as a critical phenomenon and the Flory exponent can be derived with a scaling argument.[6] Using polymer physics and Flory's theory, we have shown that protein chains behave like a real chain which is very much like an ideal chain, and that the two continuous phase transitions correspond to two processes during protein folding, namely, nucleation and condensation-collapse. The most important distinction between the two processes from the viewpoint of scaling argument is that nucleation is a 3D process and condensation-collapse a 2D process. This leads us to the following physical picture for protein folding. The folding is initiated by a few nucleation residues on various sites in a protein sequence. These nucleation residues usually have preferred local secondary structures. These local secondary structures then work as a template to guide the formation of tertiary structures as well as the further formation of secondary structures dictated by the tertiary interactions.

Our folding theory has so far been strongly supported by statistical analysis of several protein structure databases and by the statistical theory of polymer physics. Since the secondary structures of the nucleation residues can be predicted with a high probability, it is possible that some sequence patterns correspond to some secondary structures with a high degree of certainty. Our search for such sequence patterns has an advantage over the previous attempts in that the effects of the short-range and long-range interactions could be distinguished and separated clearly. This enables us to avoid many problems that might arise because the long-range interactions often make the sequence patterns not only more complicated but also very difficult to find.

Here we will describe a clustering analysis of the nucleation residues and show that the resulting clusters can be represented by a set of central sequences and constructed into a set of structural codes for the protein secondary structures. By visually analysing the characteristic physicochemical interactions and structural features of the clustered structures, we conclude that strong and clear sequence patterns in protein are often associated with the idealized hydrogen bonding patterns of the backbone geometry. Real protein sequences and structures can be considered as deviations from the ideality that is modulated by tertiary interactions among side chains and backbones that differ from protein to protein. Our findings have great significance for understanding the dynamics of protein folding, the structure prediction of protein, and the design of novel functional proteins.

2. Materials and methods

All the protein sequences and structures are based on available databases. Our method of cluster analysis of the nucleation sequences and statistical analysis of the sequence-structure relationship in the proposed protein structural codes is summarized in a flow chart shown in Fig.1.

A. Protein structure database

The structure database used in this work was extracted from SCOP-ASTRAL version 1.63 (URL: astral.berkeley.edu).[7,8] It consisted of 762 unique fold classes found in the Protein Data Bank and all the protein chains were included in our analysis whether they were determined by x-ray or by NMR and whether there were chain breaks or not.

B. Secondary structure prediction

Since we want to extract the nucleation residues and their flanking sequences, we need to first make the secondary structure prediction with the method we have developed. The details of the method have been described in Refs.[4, 5]. The nucleation residues are defined by selecting the residues whose prediction accuracy or sequence conservation is the highest (20%). Statistically this is equivalent to selecting the residues with the highest reliability score for prediction (z-reliability score). The reliability score as defined has been shown to be directly correlated with the conser-

vation of the sequence. In making the prediction for all the proteins in the database, the similarity matrix BLOSUM62 and the PSI-BLAST profile method[4,5,9] have been used. For each nucleation residue in a protein, the flanking sequences of a window size of ± 8 residues were extracted to make a sequence segment of 17 residues, with the centre residue being the nucleation residue. These sequence segments were the raw data for the clustering analysis in the present work.

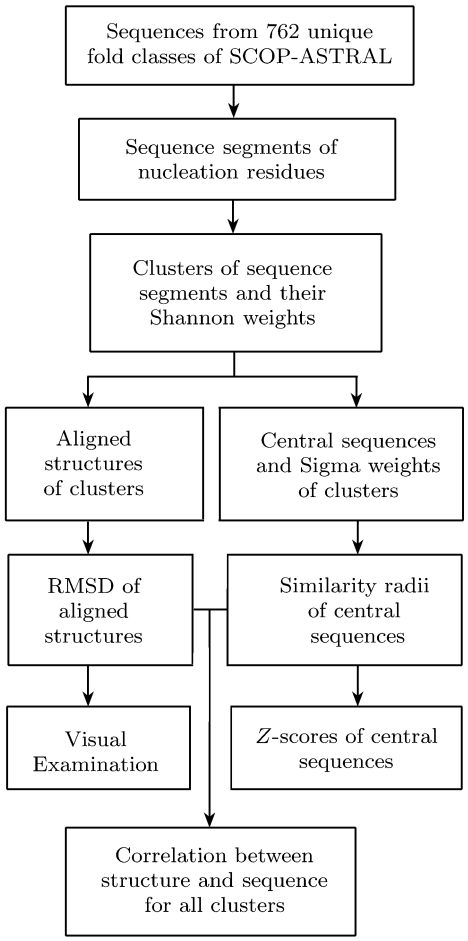


Fig.1. Flow chart of the procedure for the clustering process and the sequence-structure relationship analysis.

C. Clustering of the nucleation residues

Clustering analysis was done using the single-linkage algorithm with program OC.[10] The similarity measure between sequence segments is defined as the sequence similarity between two aligned sequences without any gap. Since all sequence segments were of the same length, it was easy and fast to calculate the sequence similarity. We also tested cluster lengths of 21 residues, but found 17 to yield the most reliable results (although this set a limit on the correlation length of helices, as shown below).

For each cluster we calculate a cluster similarity (the diameter of the cluster). Since we use the single-linkage algorithm to divide clusters, the cluster similarity given by program OC is the minimal similarity between any pair of member sequences within the given cluster. When the cluster similarity is small, the cluster will contain many sequences and the cluster size is usually large. On the other hand, when the cluster similarity is large, the cluster will contain very few sequences with a minimum cluster size of only two sequences. Since we are only interested in nontrivial clusters with medium sizes, we apply a lower and upper bound to the size of the clusters to be included for structural code analysis. Considering our sample size, that is, the number of known structures and the corresponding number of nucleation sequence segments extracted, we expect the optimal size of the clusters to be between 15 and 100 sequences. Two empirical similarity cut-offs were also applied so that the clusters to be considered for structural code analysis will not contain such unreasonable clusters that satisfy size criteria but have unusual low cluster similarity (that is, they are sparsely distributed clusters). The cutoffs used were -2.2 and -1.0. In fact, these two sets of cutoffs have similar effects and are consistent with each other.

For each cluster we construct an occurrence frequency table for amino acid usage. For each residue position relative to the centre residue, we obtain the profile of the occurrence frequency of amino acids based on the member sequences of a cluster. After normalization, we calculate the entropy (the negative of the Shannon information) of the profile at each residue position. Taking the profile with the uniform distribution as the reference for maximum entropy, we calculate the relative entropy, equal to the maximum entropy minus the profile entropy. Note that the maximum entropy depends on the number of sequences in the cluster, especially when it is less than twenty. Thus, we use random sampling to calculate the maximum entropy instead of an analytical formula. After calculating the relative entropy, we use them as a set of linear weights in the calculation of the sequence similarity so that the residue positions containing more information in determining the local secondary structures are weighted more. This set of weights for each cluster will be called the Shannon weight.

D. Representation of the sequence pattern of a cluster by the central sequence

The normalized position-sensitive occurrence frequency table and weights represent an abstraction of the member sequences contained in a cluster. However, we want to further simplify the representation of this data and achieve more abstraction. Since we have postulated that we are dealing with a cluster, we could further assume that we are dealing with a good cluster whose member sequences are randomly distributed around a central sequence. This central sequence can be imagined to be the 'ideal' sequence for the cluster. Based on this idea, we define a representative central sequence for each cluster. For each residue position or column in the frequency table, we find the amino acid type that best represents the profile for that column, that is, the total weighted sequence similarity using BLOSUM62 is the highest. At the same time, we calculate a root-mean-square deviation in sequence similarity from the central sequence weighted with the Shannon weight. Higher deviation means that the frequency profile is less clustered. We use the reciprocal of the standard deviation of the sequence similarity to measure the degree of clustering for a particular type of amino acids for the corresponding column. We call this reciprocal the Sigma weight. The Sigma and Shannon weights should have similar patterns as both carry the information about the amino acid profile, but the Sigma weight also contains the information from the similarity matrix BLOSUM62 about the twenty amino acids.

E. Calculation of z-scores for the central sequences

If the central sequences are to be used as a structural code, they should be orthogonal to one another and have small redundancy. First we need to define a cluster radius for each central sequence. The cluster radius is the sequence similarity cutoff above which a sequence is considered to be a member and below which a sequence is not considered to be a member. Here the sequence similarity of a cluster is the averaged sequence similarity between a member sequence and the central sequence. For a given cluster, this cutoff value can be modulated according to what percentage of its members is included within the cluster radius around the central sequence. The cutoff percentage used was 0.2. Then, the database sequences were divided into six non-overlapping parts according to the secondary structures (helix, sheet, and coil) and whether or not they belonged to the local (nucleation) or non-local (non-nucleation) sequences. The central sequences by definition belonged to the local databases. So for each central sequence, we can find its members in the six non-overlapping databases and use the number of the members to calculate a z-score relative to a reference generated from random sequences. From the relative ratio of the six z-scores, we can estimate the redundancy between the central sequences, that is, the structural codes.

F. Analysis of the cluster profiles

The cluster profiles, namely, the normalized occurrence frequency tables of the amino acid usage, contain the important information about the sequence patterns associated with each type of secondary structures. For example, we calculated the accumulated statistics of the centre residue for helix, sheet, and coil, respectively. This type of statistics can be done for all member sequences in each cluster, for all clusters of each secondary structure, and for all sequence segments in either the local or non-local database. The centre position can be compared with its neighbouring positions. The cluster profiles are similar to the propensity table of secondary structures.[11]

G. Analysis of the aligned structures of the member sequences of the clusters

For each member sequence in a cluster, we extract its atomic structure from its corresponding PDB entry. Then we use the member sequence with the highest reliability score (i.e. the most conserved sequence) as the reference structure and superimpose the rest of the member structures onto the reference structure. We use the centre five alpha carbons of the polypeptide chain in the superposition. After the superposition, we calculate the average root-mean-square deviation (RMSD) of the member structures relative to the reference structure. The reciprocal of a cluster radius reflects the degree of the sequence convergence of the cluster and the RMSD the degree of the structural convergence. If the two values are correlated with each other, it suggests a possible specific sequence-structure correlation between each specific member sequence and the respective member structure.

H. Visual examination of the aligned structures

The superimposed structures have been inspected visually with PyMol.[12] All the member structures can be displayed at the same time or each at a time. Hydrogen bonding patterns, especially the backbone hydrogen bonds, can be displayed individually. Amino acid residues can be coloured according to their types.

Using visual inspection, we examined the characteristic structural features and characteristic stabilizing physicochemical interactions of the proposed structural codes. Our rationale is that any sequence pattern having characteristic structure features should have the corresponding characteristic stabilizing interactions in order to become a structural code.

I. The prediction of the identity of the structural code for a protein sequence

In z-score calculation we usually need to establish a background noise level. There are many ways to do this. We usually use the simplest way possible, that is, by generating random sequences from the twenty amino acids. In this case, a structural code is the central sequence of a cluster. We slide this code sequence through the protein sequence and calculate the sequence similarity for a given window size. The corresponding background noise for each code sequence can be obtained by generating random sequences and calculating an average sequence similarity and its standard deviation by sliding the code sequence through the random sequences. Then we can use the average and the standard deviations to calculate the z-scores for the structural codes. By comparing the z-scores we can select the most probable structure code for each residue in a protein.

3. Results

A. Central sequences with the position-sensitive weights as structural codes

There were 762 PDB entries in the unique fold classes of proteins in the SCOP-ASTRAL database version 1.63.[7−8] Using DSSP[13] we assigned each residue to one of the three states: helix, sheet, and coil. As a result, there were totally 50959, 27534, and 54692 sequence segments extracted from these proteins for the helix, sheet, and coil secondary structure state, respectively. By predicting the secondary structures with program S2CHE,[4] we selected for each protein the nucleation residues whose reliability z-scores were the highest (20%). So there were 9915, 5175, and 6807 nucleation residues for helix, sheet, and coil, respectively. These nucleation residues were clustered, resulting in 155, 80, and 107 clusters for helix, sheet, and coil, respectively. Then we calculated the position-sensitive Shannon weights for each cluster. (The detailed results are available from FJ upon request.) In Fig.2, we see that the Shannon weights decrease gradually as the residue position goes away from the centre position, with the helix falling off most slowly, the sheet medium slowly, and the coil most rapidly. This indicates that helix has the longest correlation range, the sheet a bit shorter, and the coil the shortest. Therefore, we estimated the correlation ranges for helix, sheet, and coil to be about 17 (limited by the cutoff), 9, and 5 residues, respectively, similar to the previous finding.[14] We also noticed from Fig.2 that Shannon weights for helix had a pattern alternating between high and low. Another interesting point is that the highest weight did not occur at the centre position. Finally, we noticed that coil had the highest Shannon weight, which suggests a stronger sequence convergence at this residue position.

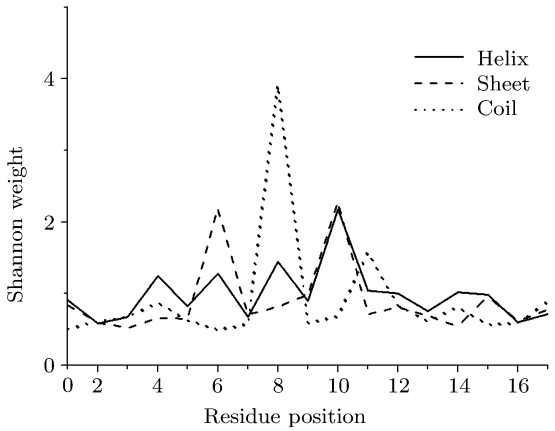


Fig.2. Shannon weights for helix, sheet and coil. The three lines are the Shannon weights for the first cluster of helix, the first cluster of sheet, and the fourth cluster of coil, respectively.

For each cluster we have derived a central sequence that is most similar to the weighted average 'sequence' of the member sequences of the cluster and used it to represent the sequence pattern of the cluster. Since the contribution from different residue positions to the sequence pattern depends on either the Shannon weight or the Sigma weight, they are an integral part of the structural codes. (Both the central sequences and the corresponding weights for all the clusters are available from FJ upon request.)

Secondly, it was found that the redundancy between the structural codes of different secondary structure states but belonging to the same type of database (local only) was very small, estimated to be well below 1%. The redundancy between the structural codes of the same secondary structure state but

belonging to different types of databases (local or nonlocal) was also very low, about 3%, 5%, and 7% for helix, sheet, and coil, respectively. This is because there are four times as many sequences in the nonlocal database as in the local database.

Thirdly, the amino acid usage in the structural codes was compared among helix, sheet, and coil. In Table 1, the amino acids are listed from top to bottom according to descending occurrence frequency. The group of amino acids at the top of the columns are usually the formers and those at the bottom the breakers for the respective secondary structures. The average weight associated with each amino acid was also calculated and shown in Table 1. The higher the either Shannon or Sigma weight is, the more contribution the corresponding amino acid makes in determining the local structure of a central sequence. For helix, E, L, A, K and R are frequent and have high weight so they are clearly formers, while C, P and G are not very frequent but have high weight and thus they are breakers. For sheet, V, I, L, T and F are important formers. For coil, G, P, and D are important formers. N and S are also formers for coil but with lower weights. Table 1 contains similar information to the propensity table of helix, sheet, and turn,[11] and the formers and breakers identified here are similar. The difference is that the propensity table does not have the weight to measure how specific an amino acid is in determining the secondary structures. To look for specific sequence patterns we need the weights for all amino acids. Thus it is always helpful to know both the occurrence frequency and the weight of an amino acid when one wants to estimate its structural effect on helix, sheet, and coil.

Table 1. Amino acid usage of the structural codes

Helix	Shannon	Sigma	Sheet	Shannon	Sigma	Coil	Shannon	Sigma
E	1.19	0.68	V	1.11	0.62	G	2.08	1.34
L	1.41	0.73	I	1.60	0.71	P	2.07	1.41
A	0.97	0.57	L	1.37	0.69	D	1.28	0.84
K	1.07	0.62	T	0.78	0.50	N	0.97	0.60
R	1.20	0.75	F	1.23	0.77	S	0.72	0.31
Q	0.89	0.48	Y	0.85	0.74	E	0.73	0.50
I	1.46	0.63	A	0.78	0.35	K	0.73	0.52
V	1.08	0.48	S	0.66	0.29	T	0.69	0.47
D	1.51	1.04	R	0.48	0.61	A	0.72	0.62
F	1.30	0.81	K	0.82	0.56	R	0.48	0.60
S	0.70	0.35	C	3.42	2.44	L	0.84	0.66
N	0.88	0.58	M	0.82	0.56	Q	0.62	0.40
M	0.79	0.34	W	1.52	1.65	H	0.59	0.89
Y	0.88	0.67	E	0.38	0.43	V	0.70	0.46
T	0.73	0.43	Q	0.56	0.31	F	1.22	0.77
H	0.86	0.86	N	0.64	0.59	I	1.37	0.69
W	2.01	1.89	H	0.86	0.92	C	0.95	1.15
C	1.61	1.47	G	1.86	1.32	Y	0.40	0.87
P	1.07	1.00	D	1.08	0.80	M	0.91	0.41
G	1.86	1.27	P	1.12	1.21	W	1.52	1.65

Note:Amino acids are listed from top to bottom in the descending order according to their occurrence frequency. The average Shannon and Sigma weights for each amino acid are averaged over the corresponding central sequences with window sizes of 9, 5, and 3 for helix, sheet, and coil, respectively.

Finally we look at the specific sequence patterns in helix, sheet, and coil. For helix, the most salient feature is the alternating polar–nonpolar pattern with the same polarity on residues i and $i+4$, which reflects the sequence pattern of the helix-wheel structure[15] and its characteristic amphipathy. Other not so regular sequence patterns are probably due to deviations from this ideal arrangement. For sheet, G has the highest Shannon weight, 1.86, without C considered, while V and I are the most frequent, as shown in Table 1. There is at least one G in 70 out of 80 central sequences. As confirmed by visual inspection, G is

most often located in beta-turns. Moreover, for sheet, amino acid V or I is often seen at the centre of a hydrophobic interaction consisting of three nonpolar residues along a beta-strand, corresponding to consecutive numbers of nonpolar residues every, or every other residue in the central sequences. For coil, the most important amino acids are G, P, D and N with the Shannon weights of 2.08, 2.07, 1.28, and 0.97, respectively. Within two residues from the centre, out of 107 central sequences, 74 have G and 61 have P. They almost all occur in beta-turns of type I or II.[16] Otherwise, each central sequence has its own particular pattern best represented by its sequences and weights.

In addition, when Cys is present in a coil, we found that it is most likely involved in a disulfide bond in a loop or a beta-strand. Table 1 also shows that the most acidic amino acid, Asp, is among the most frequent in coil. Using the 'T' state as turn, 'H' as helix, and 'E' as sheet defined by DSSP,[13] we found that there are 2764, 786, 1026, and 1052 pairs of residues for HT, TH, ET, and TE conformations, respectively. There are nearly four times as many Helix-Turn as there are Turn-Helix, which is consistent with our expectation that a negatively charged turn should usually be located at the C-terminus of helix due to its favourable interaction with the helix dipole.[17−19]

B. Characteristic structural features and stabilizing interactions

We have calculated the correlation coefficient between the reciprocal of the cluster radius of the member sequences and the average RMSD of the member structures for all clusters in helix, sheet, and coil. For a window size of 5 (±2 residues about the centre residue), the correlation coefficients are −0.06, −0.20, and 0.65 for the clusters of helix, sheet, and coil, respectively. Larger window sizes gave smaller correlation coefficients but had the same trend. Thus it was concluded that there was almost no specific sequence-structure correlation for helix and sheet, but there was a strong specific correlation for coil. In other words, more convergent sequences of helix or sheet do not necessarily lead to more convergent structures than just being an average helix or sheet. In fact, the average structural similarities of helix and sheet are slightly higher than that of coil, but the sequence variations of helix and sheet are much larger than that of coil. Therefore, only coil shows a strong specific sequence-structure correlation, that is, more convergent sequences led to more convergent structures.

Visual examination of the clustered structures shows that the structural codes of helix and sheet consist of typical structures with standard idealized geometry (Fig.3). The most frequent structure for helix in the clusters is the amphipathic helix-wheel structure (Fig.3a),[15] in which there is a continuous patch of nonpolar residues along the helical axis. Other structures have various deviations from this ideal geometry, in which there are bends or partial unwinding of helices. But nonpolar residues always try to form a continuous patch. Polar residues seem to do so but with less certainty. It is known that the specific charge and polar interactions, the helical dipole interaction, and the hydrophobic interaction all play a crucial role in stabilizing helix.[11,17−19] Here the sequence patterns suggest that the hydrophobic interaction is the most important stabilizing interaction and that this interaction should have a significant effect on many-body interaction.

We have identified three common structural elements and the corresponding stabilizing interactions from the structures clustered for sheet: They are the beta-turn, the across strand interaction, and the within strand interaction. We are relatively familiar with the first and second types of stabilizing interactions. They usually consist of hydrogen bonding, polar interaction or salt-bridge, and hydrophobic interaction. But the third type of stabilizing interaction has not yet been described before. We propose three reasons for its existence. One is based on visual inspection of sequences and structures. We have seen many cases of three aliphatic residues aligned together near the centre residue in beta-sheets. Their side-chain atoms (mainly the gamma carbon atoms) are close enough to make physical contact with each other. We also notice that the fork of the branched chains of Val, Ile, and Leu tends to lie in the plane parallel with the alpha carbon trace of the beta-strand (Fig.3b). The second reason is based on the statistics shown in Table 1. In sheet, Val and Ile are more frequently present than Leu. We believe the reason for preferring Ile than Leu is because the branched chain is more rigid at the fork and the beta-sheets are taking advantage of this rigidity to enhance the nonpolar interactions within strand. The third reason comes from a previous observation that some amino acid sequences have intrinsic preference for beta-conformations.[20] Consequently, we conclude that a beta-sheet is stabilized by multiple factors involving three types of stabilizing interactions and structural elements: turn, across-

strand interaction, and within-strand interaction.

Since there is a strong correlation between the central sequences of coil and their structures, each cluster should converge to a particular structure. We have found that most member structures converge to the standard, idealized beta-turn of type I or II. There are some loops but they are more like turns with extra residues inserted. Loops often contain multiple Ser or Gly in the central sequence. The common structural features of the turns in the clusters are similar to that of the ideal beta-turn. There are a 1–4 hydrogen bond from residue $i + 1$-NH to $i + 4$-C=O[20] and at least one polar residue in the turn. The hydrogen bond constrains the backbone geometry of the turn, and the polar residue, often exposed to the solvent, helps to stabilize the turn. Further away from the centre residue there are often additional polar residues that further stabilize the turn. Since beta-turn often occurs in beta-sheet, there is also across strand hydrophobic interaction close to the turn.

Then we examine more closely the representative member structures of cluster 6 of coil. For cluster 6 the centre residue is almost always Asp and the next C-terminal residue Gly. There are nineteen member structures and most of them are in beta-sheet conformation. Eleven out of the nineteen cluster members have the regular 1–4 hydrogen bonding[21] and the hydrophobic stabilization near the turn along the beta-sheet (Fig.4a). In the same cluster we also see two structures (Fig.4b) which have the regular 1–4 hydrogen bonding but not the hydrophobic stabilization, and, as a result, the beta-sheet has been opened up with the two strands departing from each other. In the third group of member structures, the expected regular 1–4 hydrogen bonding is missing (Fig.4c). It is substituted by the regular 1–3 hydrogen bond, which becomes the beginning of the regular hydrogen bond ladder of the neighbouring beta-sheet. There is also a hydrogen bond formed by Arg from across the strand to the carbonyl oxygen of Gly in the turn in two of the three structures of this group, which further stabilizes this type of geometry. We consider this group of structures a variation of the beta-turn, in which there is a shift of residue registration in the hydrogen bonding from 1–4 to 1–3. Interestingly, as a result of this shift, the expected across-strand hydrophobic stabilization has been preserved. In the fourth group, consisting of three member structures, both the regular 1–4 hydrogen bonding and the hydrophobic stabilization are missing (Fig.4d). The stabilization of the structures is realized through multiple hydrogen bonds from the backbone atoms to a nearby polar residue, Thr in all three cases. The geometry of these three structures is rather irregular compared to groups 1 to 3, indicating that the turn has become a loop. This type of deviations of the structures from the ideal regular structure is likely correlated with the deviations of the sequence from the ideal central sequences. It might also be influenced by the complicated non-local interactions far away from the centre residue. By analysing the four groups of the member structures with different geometries and stabilization interactions, we can conclude that the central sequence of a cluster is likely to correspond to an ideal structure, and the variations of sequences and structures can be considered as deviations from the ideal sequence-structure mapping relationship, as represented by the structural code.

4. Discussion

A. The structural codes characterized by an 'ideal' backbone geometry

We start with a nucleation hypothesis for the mechanism of protein folding.[4−5] In this work, we derived a set of structural codes for helix, sheet, and coil by clustering the nucleation residues. We show that these structural codes are well separated and independent of each other and that the representation of the sequence-structure relationship in protein can be simplified by the central sequences with the position-sensitive Shannon or Sigma weights.

The sequence patterns we find in the structural codes for helix are rather non-specific, characterized only by oscillating polar–nonpolar sequences every other residue, which is consistent with the previous experimental result.[22] This sequence pattern results from the ideal backbone geometry of an amphipathic helix wheel.[15] However, this particular geometry does not have much constraint on the sequence of helix as long as there is a continuous patch of hydrophobic residues along the helical axis. In addition, intrinsic structural similarities between aliphatic amino acids as well as between polar ones make each residue in helix easily mutable as shown in Table 2. Therefore, it is not surprising that the sequence patterns of helix are subject to variation and thus become very non-specific.

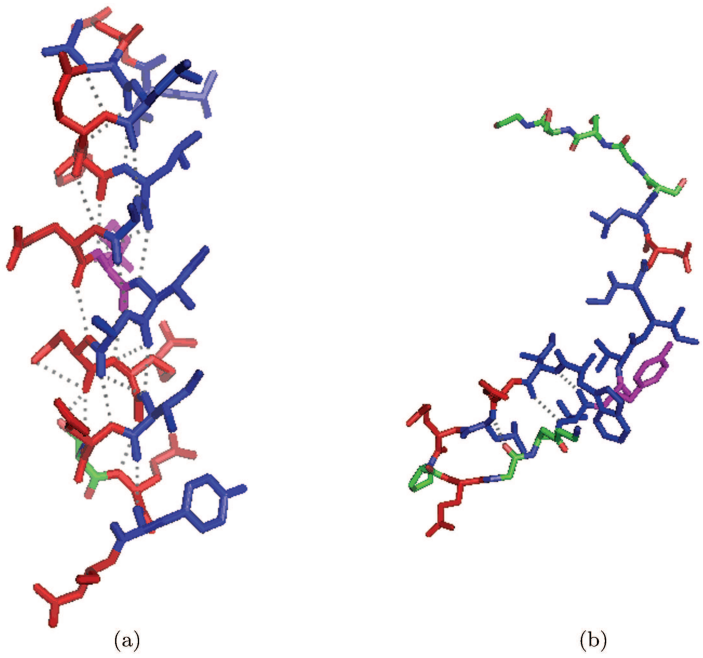


Fig.3. An amphipathic helix and a double-strand sheet in stereo plot. (a) an ideal helix having an amphipathic character with a connected, straight patch of nonpolar residues along the helical axis. (b) a typical sheet having three common structure elements: beta-turn, across-strand hydrogen bonding and hydrophobic interaction, and within-strand hydrophobic interaction. The colour scheme is that nonpolar residues are in blue, including Leu, Ile, Val, Ala, Trp, Phe, Tyr, and Met. Polar residues are in red, including His, Asp, Glu, Asn, Gln, and Arg. Other residues are in normal CPK colour, including Gly, Pro, Lys, Ser, Thr, and Cys. The centre residue is marked in pink.

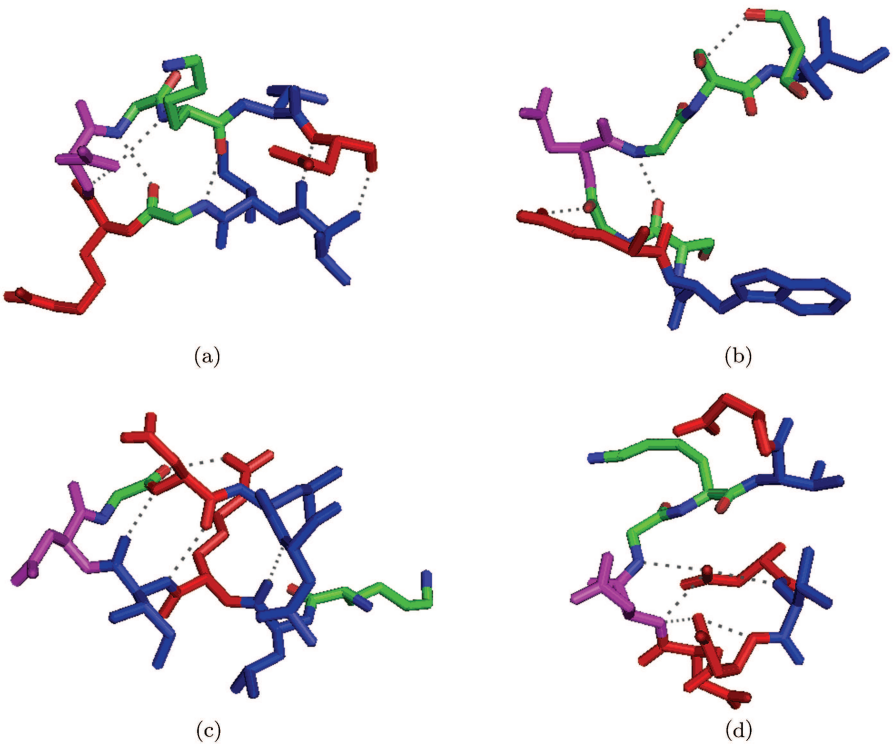


Fig.4. Clustered structures of a coil. There are 19 member structures in cluster 6 of coil, which can be divided into four groups. (a) Group 1 has both the 1–4 hydrogen bond[21] and the local hydrophobic stabilization. (b) Group 2 has only the 1–4 hydrogen bond but not the hydrophobic stabilization. (c) Group 3 does not have the 1–4 hydrogen bond but it is replaced by the 1–3 hydrogen bond.[21] (d) Group 4 shows more irregularity in which the 1–4 hydrogen bond is replaced by multiple hydrogen bonds from the side chain of a Thr nearby the backbone polar atoms. See Fig.3 for the colour scheme.

Table 2. Conserved and neutral mutations for the twenty amino acids.

Amino acid[a]	Conserved mutation[b]	Neutral mutation[c]	No. of possible mutation[d]	Conservability[e]	Mutability[f]
L	L4,I2,M2,V1	F	4	0.44	9.1
I	I4,V3,L2,M1	F	4	0.40	10.0
V	V4,I3,L1,M1	A,T	5	0.44	11.4
A	A4,S1	C,G,T,V	5	0.80	6.2
W	W11,Y2,F1		2	0.79	2.5
F	F6,Y3,W1	I,L,M	5	0.60	8.3
Y	Y7,F3,H2,W2		3	0.50	6.0
M	M5,L2,I1,V1	F,Q	5	0.56	8.9
C	C9	A	1	1.00	1.0
T	T5,S1	A,N,V	4	0.83	4.8
S	S4,A1,N1,T1	E,G,K,Q	6	0.57	10.5
G	G6	A,N,S	3	1.00	3.0
P	P7		0	1.00	0.0
H	H8,Y2,N1	E,Q,R	5	0.73	6.8
D	D6,E2,N1	Q,S	4	0.67	6.0
E	E5,D2,Q2,K1	H,N,R,S	7	0.50	14.0
N	N6,D1,H1,S1	E,G,K,Q,R,T	9	0.67	13.4
Q	Q5,E2,K1,R1	D,H,M,N,S	8	0.56	14.3
R	R5,K2,Q1	E,H,N	5	0.63	7.9
K	K5,R2,E1,Q1	N,S	5	0.56	8.9

Note: This table is a simplified version of the similarity matrix BLOSUM62. The conservability and mutability together show nicely the properties of the amino acids and how they are related to each other and how their detailed structures can make important difference.

a. The twenty amino acids are listed from top to bottom according to the order of the descending hydrophobicities.
b. Any mutation with a positive similarity score is considered to be a conserved mutation. The column format is amino acid in single letter with a similarity score attached to each letter. The first letter is the self-mutation.
c. Any mutation with a similarity score of zero is considered to be a neutral mutation.
d. The number of possible mutations is the sum of the numbers of conserved and neutral mutations excluding the self-mutation.
e. The conservability is defined as the self-similarity score divided by the sum of the similarity scores of all possible conserved mutations including the self. The conservability of an amino acid measures its ability to resist change to other amino acids.
f. The mutability is defined as the number of possible mutations divided by the conservability. The mutability measures how easy it is for an amino acid to change to other amino acids with similar structures and properties.

For sheet, we have identified three important structural elements and the associated stabilizing interactions: turn, across-strand interaction, and within-strand interaction. All of them contribute to the formation and stabilization of sheet. Due to the involvement of the hydrophobic interaction in both across-strand and within-strand interactions, the sequence patterns for sheet are also rather non-specific for the same reason as that for helix, that is, there are the lack of constraint on sequence and the intrinsic mutability of hydrophobic amino acids.

For coil, we have found a very specific sequence-structure relationship. The structures clustered in coil are in fact mostly beta-turn. Random coil or loop occurs in structural codes only when there are more than one Gly, Pro, or Ser in the sequence pattern. We have found that the specific structures to which the clustered sequences converge are most strikingly characterized by the ideal, regular specific hydrogen bonding patterns. For coil, the identified hydrogen bonds correspond to the structural patterns of the beta-turn of type I or II.

The above results have led us to the conclusion that protein structures with an ideal geometry have the most specific and conserved sequence patterns, and are the most stable. These specific sequence pat-

terns should be the ideal candidates for the structural codes we are looking for. By ideal geometry we are particularly referring to the ideal backbone geometry and hydrogen bonding patterns in helix, sheet, and turn.

Therefore, our view of the folding principle is that ideal sequences tend to be folded into ideal structures (as in structural codes) which are generally more stable and therefore they should also fold first because they are generally controlled by short-range interactions. Tertiary structures resulting from long-range interactions form later and in general should have less ideal geometries. Real protein sequences and structures can be considered as 'frustrated' variations based on the ideal structural codes. In a real protein, when an unfavorable interaction or a loss of a favourable interaction occurs, there should be a compensating interaction that is either a new favourable interaction or a loss of another unfavorable interaction. Through our observation on the known protein structures, the complexity of the interactions and the corresponding structures resulting from such compensations increases in the following order: (1) an interaction is replaced by another interaction of the same type; (2) there is a small shift in residue registration but the same type of interaction is preserved; (3) an interaction is replaced by more than one new but weaker interaction of a similar type. The complexity increases particularly as more side chain atoms are involved in the hydrogen bonding, in the following order: (i) a backbone hydrogen bond is replaced by a combination of new backbone hydrogen bonds that are less ideal in geometry and weaker in strength, with a side chain hydrogen bond; (ii) a backbone hydrogen bond is replaced by multiple side chain hydrogen bonds. Tertiary interactions will in general distort the structures from ideality, but in most proteins the trace of the structural codes should not be completely wiped out. We believe that this results from the biological evolution and from the chemical structure of the twenty amino acids and the polypeptide chain.

B. The implications for protein structure prediction and design

We have simplified the similarity matrix BLOSUM62 in a very intuitive and useful way by using a table representation (Table 2). We focus on the conserved and neutral mutations for each amino acid by determining the conservability and mutability. From this table we can see L, I, V and M are mutable to each other as a group, because their side chain structures are very similar. On the other hand, P is conservative and not mutable while C and G are conservative but still mutable. We also notice that S but not T is very mutable, polar amino acids E, N and Q are the most mutable, and aromatic amino acid F is more mutable than W and Y because it can be changed into L, I and M through neutral mutation. Whenever a mutable residue occurs in a hypothetical structural code it will blur the underlying sequence pattern. Since most polar and nonpolar amino acids are mutable, most sequence patterns tend to be just polar and nonpolar. Finally, we need to remember that for structural codes, both their member sequences and member structures need to be orthogonal to one another. This explains why it is difficult to construct sequence-structure codes in general, and for helix and sheet in particular.

Table 1 and 2 together should be useful not only for analysing protein structure but also for prediction and design. For example, Glu and Leu are most frequent in helix while Val and Ile in sheet. This is understandable in structural terms. Helix has a Christmas tree structure so its side chains need to be longer to avoid steric collision with the backbone in order to interact with each other on the surface. On the other hand, the side chains in sheet are perpendicular to the strand backbone, so shorter amino acids such as Val and Ile are selected more in order to make favourable contacts. This is because the interaction Centre of a side chain should not be too far away from the backbone so as to exert sufficient influence on the backbone geometry. In general, we only need to focus on the formers of helix, sheet, and turn, located in the top rows of Table 1, while the breakers of helix and sheet happen to be the formers of turn, for example, Gly and Pro.

From the above, we see that tertiary interactions are important for the protein structures. Because tertiary interactions involve a complicated geometry, accurate calculation of their free energies is essential for determining the tertiary structures involved. This is known to be a difficult task, because the time scales reachable with the current simulation techniques is still limited and because many-body interactions can play a significant role.[23] However, our nucleation theory suggests that the folding process of a protein can be divided into short segments, which is manageable with MD simulations.[24,25] Then by combining these short segments, a composite picture of the whole

folding process can be reconstructed at both longer time scales and correlation length scales. Nevertheless, more studies are needed to make prediction and design with more accuracy. There have been many fruitful and successful attempts in the past.[26−29] Our new folding theory and the proposed structural codes should be helpful to developing new methods for theoretical calculations in the future.

C. The connection between protein folding and Flory's theory

If an unfolded protein chain in solution is like a nearly ideal chain, it is the theta-point.[6,30] The theta-point occurs when, under an appropriate solution condition, there is a perfect balance between the repulsive interaction and the attractive interaction among monomers, and the effective interaction approaches zero. Since proteins are, in fact, weakly attractive on average in the folded state, the total interaction energy should be weakly attractive. This explains why proteins in general have low stability. More evidence supporting our view comes from the fact that the internal friction of protein chains is very low as shown by recent experiments.[31,32] Experimentally the molten globular state of protein should most likely resemble the condition of the theta-point. The reason that proteins could achieve ideal chain-like behaviour is perhaps easy to understand. This is because proteins are heteropolymer with its monomers (residues) selected from twenty different types of amino acids. As a result, it is easier for a heteropolymer to achieve the perfect balance of repulsive and attractive forces than a homopolymer. Therefore it is not surprising that protein chains are made of a few hundred residues instead of a few thousand monomers as in synthetic homopolymers.

Many experiments show that there are glass transitions in protein folding, which seems to contradict the nucleation theory. But if the effect of the reduction of dimension from 3D to 2D during protein folding is taken into account, it may be possible to explain most of the experimental observations. This is because the dynamics in 2D may appear to be trapped and glass-like in 3D. When the moving chain of protein detaches itself from the template formed by the nucleation residues or the template has been destroyed by complete unfolding, for example, by heating, the dynamics will become normal again in 3D. But things can become complicated when the glass transition temperature of the condensation-collapse process is equal to the melting temperature of the nucleation process. It is often difficult to distinguish the two processes experimentally and that is why the nucleation process has not been observed for so many two-state folding proteins. By identifying structural codes it may become possible to separate the two processes with mutagenesis or modification of amino acid residues in a protein to further test our nucleation theory.

5. Concluding remarks

We have seen that nucleation is the basis on which we develop the nucleation folding theory. Structural codes in protein sequence may play a role of seeds in the nucleation process of protein folding. That is why we have derived the structural codes from the nucleation residues. Because the theory suggests non-random pathways and more or less stepwise and ordered processes for folding, we should be more optimistic about predicting the protein structure based on sequence alone and designing novel functional proteins tailored to specific needs.

The programs for clustering analysis and for predicting the identity of the structural codes for each residue in a given protein sequence have been written in Perl script. The programs and the accompanying data are available upon request to FJ (fjiang@aphy.iphy.ac.cn).

References

[1] Anfinsen C B 1973 *Science* **181** 223
[2] Levinthal C 1968 *J. Chim. Phys.* **65** 44
[3] Garnier J and Levin J M 1991 *Comput. Appl. Biosci.* **7** 133
[4] Jiang F 2003 *Protein Eng.* **16** 651
[5] Jiang F 2005 *Sci. & Technol. Adv. Mater.* **6** 860
[6] de Gennes P G 1979 *Scaling Concepts in Polymer Physics* (Ithaca, NY and London: Cornell University Press)
[7] Murzin A G, Brenner S E, Hubbard T and Chothia C 1995 *J. Mol. Biol.* **247** 536
[8] Lo Conte L, Brenner S E, Hubbard T J P, Chothia C and Murzin A 2002 *Nucl. Acid Res.* **30** 264
[9] Altschul S F, Madden T L, Schaffer A A, Zhang J, Zhang Z, Miller W and Lipman D 1997 *Nucleic Acids Res.* **25** 3389
[10] Barton G J 1993 *A Cluster Analysis Program* (University of Dundee, Scotland)
[11] Chou P Y and Fasman G D 1974 *Biochemistry* **13** 222

[12] PyMol 2004 DeLano Scientific LLC.
[13] Kabsch W and Sander C 1983 *Biopolymers* **22** 2577
[14] Baldi P, Brunak S, Frasconi P, Soda G and Pollastri G 1999 *Bioinformatics* **15** 937
[15] Stryer L 1998 *Biochemistry* 4th ed. (New York: Freeman)
[16] Nelson D L and Cox M M 2005 *Lehninger Principles of Biochemistry* 3rd ed. (W.H. Freeman and Company, New York)
[17] Shoemaker K R, Kim P S, York E J, Stewart J M and Baldwin R L 1987 *Nature* **326** 563
[18] Sali D, Bycroft M and Fersht A R 1988 *Nature* **335** 740
[19] Serrano L and Ferhst A R 1989 *Nature* **342** 296
[20] Swindells M B, MacArthur M W and Thornton J M 1995 *Nat. Struct. Biol.* **2** 596
[21] Jeffrey G A and Saenger W 1994 *Hydrogen Bonding in Biological Structures* (Berlin: Springer)
[22] Kamtekar S, Schiffer J M, Xiong H, Babik J M and Hecht M H 1993 *Science* **262** 1680
[23] Modhaddam M S, Shimizu S and Chan H S 2005 *J. Am. Chem. Soc.* **127** 303
[24] Sheng Y B, Li J, Ma B L and Wang W 2005 *Chin. Phys.* **14** 2365
[25] Ho B K and Dill K A 2006 *PLoS Comp. Biol.* **2** 228
[26] Kuhlman B, Dantas G, Ireton G C, Varani G, Stoddard B L and Baker D 2003 *Science* **302** 1364
[27] Looger L L, Dwyer M A, Smith J J and Hellinga H W 2003 *Nature* **423** 186
[28] Park S, Yang X and Saven J G 2004 *Curr. Opin. Struct. Biol.* **14** 487
[29] Kuhlman B and Baker D 2004 *Curr. Opin. Struct. Biol.* **14** 89
[30] Flory P J 1953 *Principles of Polymer Chemistry* (Ithaca, NY: Cornell University Press)
[31] Pabit S A, Roder H and Hagen S J 2004 *Biochemistry* **43** 12532
[32] Plaxco K W and Baker D 1998 *Proc. Natl. Acad. Sci. U. S. A.* **95** 13591