The role of hydrophobic interactions in folding of β-sheets

Jiacheng Li²,¹, Xiaoliang Ma²,¹, Hongchi Zhang²,¹, Chengyu Hou²,¹, Liping Shi³, Shuai Guo³, Chenchen Liao⁴, Bing Zheng⁵, Lin Ye⁶, Lin Yang⁶,², Xiaodong He³,e

¹ National Key Laboratory of Science and Technology on Advanced Composites in Special Environments, Center for Composite Materials and Structures, Harbin Institute of Technology, Harbin 150080, China
² School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150080, China
³ Key Laboratory of Functional Inorganic Material Chemistry (Ministry of Education) and School of Chemistry and Materials Science, Heilongjiang University, Harbin 150001, P. R. China.
⁴ School of Aerospace, Mechanical and Mechatronic Engineering, The University of Sydney, NSW 2006, Australia
⁵ Shenzhen STRONG Advanced Materials Research Institute Co., Ltd, Shenzhen 518035, P. R. China.

Exploring the protein-folding problem has been a long-standing challenge in molecular biology. Protein folding is highly dependent on folding of secondary structures as the way to pave a native folding pathway. Here, we demonstrate that a feature of a large hydrophobic surface area covering most side-chains on one side or the other side of adjacent β-strands of a β-sheet is prevail in almost all experimentally determined β-sheets, indicating that folding of β-sheets is most likely triggered by multistage hydrophobic interactions among neighbored side-chains of unfolded polypeptides, enable β-sheets fold reproducibly following explicit physical folding codes in aqueous environments. β-turns often contain five types of residues characterized with relatively small exposed hydrophobic proportions of their side-chains, that is explained as these residues can block hydrophobic effect among neighbored side-chains in sequence. Temperature dependence of the folding of β-sheet is thus attributed to temperature dependence of the strength of the hydrophobicity. The hydrophobic-effect-based mechanism responsible for β-sheets folding is verified by bioinformatics analyses of thousands of results available from experiments. The folding codes in amino acid sequence that dictate formation of a β-hairpin can be deciphered through evaluating hydrophobic interaction among side-chains of an unfolded polypeptide from a β-strand-like thermodynamic metastable state.

INTRODUCTION

Protein products are the basis of life on Earth and serve nearly all the functions in the essential biochemistry of life science. Each nascent protein exists as an unfolded polypeptide or random coil when translated from a sequence of mRNA to a linear chain of residues by a ribosome. The intrinsic biological functions of a protein are expressed and determined by its native three-dimensional (3D)
structure that derives from the physical process of protein folding\(^1\), by which a polypeptide folds into its native characteristic and functional three-dimensional structure, in an expeditious and reproducible manner. Protein folding can thereby be considered the most important mechanism, principle, and motivation of biological existence, functionalization, diversity, and evolution\(^2\)–\(^4\).

Based on the complexity of protein folding, the protein-folding problem has been summarized in three unanswered questions\(^1\): (i) What is the physical folding code in the amino acid sequence that dictates the particular native 3D structure? (ii) What is the folding mechanism that enables proteins to fold so quickly? (iii) Is it possible to devise a computer algorithm to effectively predict a protein’s native structure from its amino acid sequence? Moreover, another essential question is why protein folding highly depends on the solvent (water or lipid bilayer)\(^5\) and the temperature\(^6\)? The protein folding problem was brought to light over 60 years ago. In particular, since Anfinsen shared a 1972 Nobel Prize in Chemistry for his work revealing the connection between the amino acid sequence and the native conformation\(^7\), understanding of protein sequence-structure relationships has become the most fundamental task in molecular and structural biology \(^8\).

Protein folding is one of the miracles of nature that human technology finds quite difficult to follow, due to the very large number of degrees of rotational freedom in an unfolded polypeptide chain. In the 1960s, Cyrus Levinthal pointed out that the apparent contradiction between the astronomical number of possible conformations for a protein chain and the fact that proteins can fold quickly into their native structures should be regarded as a paradox (Levinthal's paradox)\(^9\), so there must be mechanisms that allow polypeptide chains to find the native states encoded in their sequence. As stated in Anfinsen's Dogma, the well-defined native 3D structures of small globular proteins are uniquely encoded in their primary structures (the amino acid sequences), is kinetically reproducible and stable under a range of physiological conditions, and can therefore be considered as an issue of the certainty.

Many proteins or protein domains, relatively rapid and efficient refolding can be observed in vitro, thus proteins may be regarded as "folding themselves" following explicit folding pathways\(^1\). Protein folding is considered a free energy minimization or a relaxation process that is guided mainly by the following
physical forces: (i) formation of intramolecular hydrogen bonds, (ii) van der Waals interactions, (iii) electrostatic interactions, (iv) hydrophobic interactions, (v) chain entropy of protein, (vi) thermal motions\textsuperscript{1,10}. Among them, hydrophobic effect is normally thought to play a decisive role\textsuperscript{11}. Currently, the generally accepted hypothesis in the field is to conceive of protein folding in a funnel-shaped energy landscape, where every possible conformation is represented by a free energy value. The rapid folding of proteins has been attributed to random thermal motions that cause conformational changes leading energetically downhill toward the native structure corresponds to its free energy minimum under the solution conditions \textsuperscript{1,10}. However, there are both enthalpic and entropic contributions to free energy of protein that change with temperature and so give rise to heat denaturation and, in some cases, cold denaturation\textsuperscript{12}. So far the hypothesis haven’t been able to decipher the folding code and therefore aren’t generally able to read a sequence and predict what shape it will adopt.

The interaction of protein surface with the surrounding water is often referred to as protein hydration layer (also sometimes called hydration shell) and is fundamental to structural stability of protein, because non-aqueous solvents in general denature proteins \textsuperscript{13}. The hydration layer around a protein has been found to have dynamics distinct from the bulk water to a distance of 1 nm and water molecules slow down greatly when they encounter a protein\textsuperscript{14}. Thus, hydrophilic side chains of proteins are normally hydrogen bonded with surrounding water molecules in aqueous environments, thereby preventing the surface hydrophilic side-chains of proteins from randomly hydrogen bonding together \textsuperscript{14,15} \textsuperscript{16}. This is the reason why proteins usually do not aggregate or crystallize in unsaturated aqueous solutions\textsuperscript{17} , even though the solvent-facing surface of the proteins is usually composed of predominantly hydrophilic regions. Experiments have also shown that secondary structures of protein (such as $\alpha$-helices and $\beta$-sheets) are stabilized by hydrogen bonds between the N-H groups and C=O groups of the main chain\textsuperscript{18,19}. This also indicates that the shielding effect of surrounding water molecules prevent hydrophilic side-chains from interfering with the formation of secondary structures during protein folding. Thus, water molecules should be able to saturate the hydrogen bond formations of hydrophilic side-chains and the main chain before the protein folding \textsuperscript{14-16}, due to water molecules have very strong polarity\textsuperscript{20,21}. 
This is the reason why intrinsically disordered proteins (IDPs) and regions (IDRs) can make up a significant part of the proteome. Before the folding of secondary structures, the early steps of protein folding may be not directly dominated by the formation of intramolecular hydrogen bonds, due to the shielding effect of surrounding water molecules. Thus, this problem may lie in our lack of understanding of the hydrophobic interaction among neighbored side-chains of unfolded proteins at early steps of the folding, given the lack of awareness of the importance of the shielding effect of water.

Almost all experimentally determined native tertiary structures of water-soluble proteins have a hydrophobic core in which hydrophobic side-chains are buried from water. Incidentally, polar residues interact favorably with water, thus the solvent-facing surface of the peptide is usually composed of predominantly hydrophilic regions. Minimizing the number of hydrophobic side-chains exposed to water, namely, hydrophobic collapse thus has been regarded as one of the most important driving force for protein folding processes. Experimental methods such as laser temperature jumping technology and single molecule experimental techniques have revealed that protein folding first leads to the formation of secondary structures (α-helices and β-strands), and the tertiary structure is formed by the folding of secondary structures. It is likely that the nascent polypeptide forms initial secondary structure through creating localized regions of predominantly hydrophobic residues due to hydrophobic effect. The secondary structures interacts with water, thus placing thermodynamic pressures on these regions which then aggregate or "collapse" into a tertiary conformation with a hydrophobic core. Therefore, protein folding is highly dependent on folding of secondary structures as the way to hierarchically pave a native folding pathway that lead to formation of correct tertiary structures and cause conformational changes leading energetically downhill toward the native globular structure that possesses the minimum free energy. Thus, decipher of the folding codes in amino acid sequence that dictate the secondary structures formation should be regarded as a key to crack the protein folding problem. Among types of secondary structure in proteins, the β-sheet is the most prevalent. If the controlling mechanism for β-sheet folding can be revealed, it would remarkably promote solution of the protein folding problem.

Currently, several hypotheses has been proposed for explaining the folding mechanism of β-sheet. The hydrophobic zipper hypothesis indicates that a hairpin is first formed before hydrophobic contacts act as
constraints which bring other contacts into spatial proximity\textsuperscript{30}. This leads to further constrain and causes the rest of the contacts to zip up. Munoz \textit{et al} proposed that the folding of a $\beta$-hairpin initiates at the turn and propagates towards the tails\textsuperscript{31}. In particular, they found that stabilization through hydrophobic contacts between residues and hydrogen bonding interaction are important for the formation of the $\beta$-hairpin. Petrovich \textit{et al.}\textsuperscript{32} studied a 37-residue triple-stranded $\beta$-sheet protein via MD simulations. Their results indicate that a $\beta$-hairpin first appears before the third strand joins in to complete the $\beta$-sheet at the end of the folding process. They ascribe the folding mechanism of the $\beta$-sheet to a combination of initial hydrophobic collapse and zipper mechanism, which serve to nucleate the hairpin formation. Notably, all the three mechanisms above suggest that the folding of a $\beta$-sheet is necessarily preceded by the occurrence of a $\beta$-turn. We are still missing a "folding mechanism" for $\beta$-sheets. By mechanism, we mean a narrative that explains how the time evolution of a $\beta$-sheet folding development derives from its amino acid sequence and solution conditions.

\textbf{Results}

$\beta$-sheet folding highly depends on the temperature\textsuperscript{5}, where $\beta$-sheets can form in as little as 1 microsecond after the temperature jumping\textsuperscript{33-35}. $\beta$-sheets consist of $\beta$-strands connected laterally by at least three backbone hydrogen bonds, forming a generally pleated sheet. A $\beta$-strand is a stretch of polypeptide chain typically 3 or more amino acids long with backbone in an extended conformation. It most like that the $\beta$-strands exist before the folding of $\beta$-sheets. Because it is difficult to explain how the folding process of a $\beta$-sheet (i.e., laterally hydrogen bonding process of segments of unfolded polypeptide) is accompanied by stretching process of the segments of polypeptide into $\beta$-strands. There must be mechanisms that allow polypeptide chain segments to find the states of $\beta$-strands encoded in their sequence. There also must be some physical effects providing the long-range attractive force among $\beta$-strands for the $\beta$-sheets formation.

Experimental evidences of the folding of unfolded proteins provide corroboration for a hypothesis that folding initiation sites arise from hydrophobic interactions\textsuperscript{11,36}. The folding of $\beta$-strands and $\beta$-sheets may be driven by hydrophobic interactions, as the nascent polypeptide may form initial primary structure
through creating localized regions of predominantly hydrophobic residues\textsuperscript{29}. Hydrophobic effect most likely can contribute to the formation of β-sheets through multistage aggregations of neighbored hydrophobic groups of unfolded polypeptides, which lead to the formation of β-strands, and consequently fold into β-sheets. A β-sheet always is amphipathic in nature, namely, contain hydrophilic surface areas and hydrophobic surface areas. Note that the hydrophobic attraction (due to the hydrophobic effect) among adjacent side-chains on one side or the other side of a β-strand may be common in experimentally determined protein structures, which should be considered as an evidence for hydrophobic effect dominating the formation of β-strands.

It has previously been noted that many amino acid side chains contain considerable nonpolar sections, even if they also contain polar or charged groups\textsuperscript{36}. Namely, hydrophilic side-chains are not completely hydrophilic. The hydrophilicity of hydrophilic side-chains is normally expressed by C=O or N-H\textsubscript{2} groups at their ends, and the other portions of hydrophilic side-chains are hydrophobic, because the molecular structures of these portions are basically alkyl and benzene ring structures, as shown in Figure 1. Folding initiation sites of β-brands might therefore contain not only accepted “hydrophobic” amino acids, but also larger hydrophilic side-chains\textsuperscript{36}. If formation of β-brands is driven by hydrophobic interactions among neighbored side-chains of unfolded polypeptide, we should be able to find experimental evidence of the hydrophobic interaction in the Protein Data Bank (PDB) achieves, due to hundreds of thousands of β-sheet structures have been experimentally determined. In an aqueous environment, the water molecules tend to segregate around the “hydrophobic” side chains of the nascent protein, creating hydration shells of ordered water molecules\textsuperscript{37}. An ordering of water molecules around a hydrophobic region increases order in a system and therefore contributes a negative change in entropy (less entropy in the system)\textsuperscript{38}. The water molecules are fixed in these water cages which drives the hydrophobic collapse, or the aggregation of the hydrophobic groups. Thus, the hydrophobic interaction among neighbored side-chains in sequence can introduce entropy back to the system via the breaking of their water cages which frees the ordered water molecules\textsuperscript{39}. If hydrophobic interactions among neighbor side-chains in amino acid sequences provide the structural stability for β-brands formation, we must can find out that the phenomenon of a large hydrophobic surface area covering on one side or the other side of a
β-strand is prevail in almost all experimentally determined β-sheets. If the phenomenon of hydrophobic side-chains tend to cluster together on one side of adjacent β-strands of a β-sheet is prevail in almost all experimentally determined β-sheets, we may demonstrate that the hydrophobic interaction among the neighbored side-chains responsible for β-sheet-folding initiation.

The capability of an amino acid residue to get involved in the hydrophobic attraction with neighbored residues in sequence can be evaluated by the exposed alkyl and benzene ring structures of the side-chain, as shown in Fig.1, in which 20 kinds of amino acid residue are divided into four groups. Arginine-R, Histidine-H, and Lysine-K can involve in hydrophobic interaction with adjacent hydrophobic side-chains in sequence due to their long hydrophilic side chains contain long nonpolar alkyl structures, see Fig.1A. Cysteine-C, Isoleucine-I, Leucine-L, Methionine-M, Tryptophan-W, Phenyllalanine-F, Tyrosine-Y, and Valine-V can fully involve in hydrophobic interaction with adjacent side-chains due to their high hydrophobicity, see Fig.1B. Glutamate-E, Glutamine-Q, Threonine-T, and Alanine-A would allow limited participation in hydrophobic interaction with neighbored side-chains in sequence due to their exposed hydrophobic proportions is relatively small, see Fig.1C. Aspartate-D, Asparagine-N, Serine-S, Proline-P, and Glycine-G basically can’t participate in hydrophobic interaction with adjacent side-chains in sequence due to the hydrophobic proportions of their side-chains are too small or being occluded by hydrophilic groups, see Fig.1D.

A de novo designed protein with curved β-sheet (PBDID: 5TPJ) is a good example for illustrating the phenomenon of the hydrophobic attraction (due to the hydrophobic effect) among adjacent side-chains on one side of each β-strand of the protein, see Fig.2. To illustrate the hydrophobic attraction, we highlight the hydrophobic surface areas of adjacent side-chains on each β-strand of the protein based on the experimentally determined protein structure as shown in Fig. 2C and 2D. Noting that every β-strand is characterized by a large hydrophobic surface fully covering one side of the β-brand (the inner side), and caused each side-chains is parallel to every other side-chain of each strands due to the hydrophobic interaction. Parallel distribution of adjacent peptide planes of these β-strands also causes adjacent side-chains to distribute on opposite sides of the main chain and each carbonyl oxygen atom in a peptide plane tends to hydrogen bond with an amide hydrogen atom in an adjacent peptide plane due to the electrostatic
attractions between them, except the Proline-P\textsuperscript{15}. Parallel distribution of neighbored “hydrophobic” side-chains in a β-strand can effectively introduce entropy back to the system via the merging of the water cages of the side-chains which frees the ordered water molecules, see Fig.2D. Thus, β-strand should be considered as a metastable state for unfolded polypeptides corresponds to its free energy minimum under the solution conditions, creating localized regions of predominantly hydrophobic side-chains\textsuperscript{15}.

We use another small-molecule protein (PBDID:1OUR) as the example to demonstrate the role of hydrophobic interactions among neighbored side-chains played in formation of β-strands, β-turns and β-sheets, see Fig.3. The protein is mainly composed with β-strands and 10 β-turns. Every β-strand of the protein is also characterized by a large hydrophobic surface fully covering one side or the other side of the β-brand, see Fig.3A. Aspartate-D, Asparagine-N, Serine-S, Proline-P, Glycine-G most likely contribute to formation of β-turns in protein folding, due to the other neighbored side-chains in amino acid sequence tend to hydrophobic attract with each other through bypassing these residues (see Fig.1d). Thereby, Aspartate-D, Asparagine-N, Serine-S, Proline-P, Glycine-G can be classified as a hydrophobic blocking (R\textsubscript{B}) group. It is worth noting that almost all the 10 β-turns of the protein are composed with two or more residues of Aspartate-D, Asparagine-N, Serine-S, Proline-P, Glycine-G, see Fig.3A and 3B. This indicates that two or more adjacent R\textsubscript{B} residues can effectively block hydrophobic attraction among neighbored side-chains in sequence at both side of a strand. We plot the protein structure into three parts according to three segments of the amino acid sequence to illustrate the hydrophobic collapse among neighbored β-strands in sequence, see Fig.3B and 3C. Hydrophobic interactions among these β-strands may drive them collapse together through bending the unfolded polypeptide at the location of these R\textsubscript{B} residues, namely, bypassing these R\textsubscript{B} residues at the turns to achieve the hydrophobic collapse. This also indicates that hydrophobic attraction among neighbored side-chains drive the β-strands formation and then cause hydrophobic attraction among the neighbored β-strands and formation of the β-sheets, due to β-strands formation create localized regions of predominantly hydrophobic residues and place thermodynamic pressures on these regions under the solution conditions. Formation of β-sheets also make β-strands aggregate or "collapse" into a tertiary conformation with a hydrophobic core. Thereby, we speculate that folding of β-sheets is triggered by multistage hydrophobic interactions among
neighbored side-chains of unfolded polypeptides, enable β-sheets fold reproducibly following explicit physical folding codes in aqueous environments.

We use 1000 experimentally determined small protein structures to further demonstrate the hydrophobic-effect-based folding mechanism for β-sheets. All the 1000 small proteins were randomly selected from the PDB. 3235 β-strands can be identified in the 1000 protein structures by using the PDB archive and the STRIDE software. From analysis of all the 3235 β-strands of the 1000 proteins in PDB, we find out that the feature of hydrophobic attraction (due to the hydrophobic effect) among adjacent side-chains on one side or the other side of a β-strand covering the length of the β-strand is prevail in all the experimentally determined β-strands (see Supplementary S5). This indicates that the hydrophobic interaction among the neighbored side-chains responsible for the formation of β-strands.

Aspartate-D, Asparagine-N, Serine-S, Proline-P, Glycine-G can’t effectively hydrophobic attract with neighbored side-chains in sequence, see Fig.1D. Thus, Aspartate-D, Asparagine-N, Serine-S, Proline-P, Glycine-G most likely lead to β-turns formation in protein folding, due to the other neighbored side-chains in amino acid sequence tend to hydrophobic attract with each other through bypassing these residues. The β-turn is the third most important secondary structure after helices and β-strands. β-turns have been classified according to the values of the dihedral angles φ and ψ of the central residue. β-turns can be easily identified in between β-strands or α-helices of the protein structures using the PDB archive and the STRIDE software. We identified 5776 β-turns in the 1000 protein structures, include about 1780 β-hairpin turns. We find out that about 97.4% of the β-turns contain at least one Aspartate-D, Asparagine-N, Serine-S, Proline-P or Glycine-G residue, as illustrated in Supplementary 2. Whereas, most of the rest no-Rβ β-turns contain at least one Glutamate-E, Glutamine-Q, Threonine-T, and Alanine-A residue. This indicates that Glutamate-E, Glutamine-Q, Threonine-T, and Alanine-A may contribute to the formation of β-turns due to their exposed hydrophobic proportions is relatively small. Moreover, about 99.3% β-hairpin turns contain at least one Aspartate-D, Asparagine-N, Serine-S, Proline-P or Glycine-G residue, see Supplementary 2.
Two R<sub>B</sub> residues coded together normally shouldn’t be able to present at the middle of a long straight β-strand. Because the other residues of the strand at both sides of the two R<sub>B</sub> residues tend to hydrophobic aggregate together and thus would bend the strand at the two R<sub>B</sub> residues to achieve the hydrophobic interaction. However, we can still identified 29 long β-strands (each β-strands contain more than 12 residues), which are characterized by two adjacent R<sub>B</sub> residues locating at the middle of the β-strands through scanning the 1000 protein structures by using the STRIDE software<sup>42</sup>. By checking these long β-strands using PyMOL software, we find out that 24 of these long β-strands actually curved exactly at their two R<sub>B</sub> residues in the amino acid sequences, demonstrating the capability of R<sub>B</sub> residues to cause β-turns formation, see Fig. 4. The other 5 long β-strands either have three or more R<sub>B</sub> residues coded together or have R<sub>B</sub> residues located at one end of the strands that make the hydrophobic blocking region extend to the ends of these β-strands, thus undermining the hydrophobic interaction between the both ends of these β-strands, see Supplementary S3. The long β-strand of the 1YV7 protein curved at a sequence segment of threonine-threonine-terine-glutamate (TTSE), see Supplementary S3. This indicates that Glutamate-E, Glutamine-Q, Threonine-T, and Alanine-A may also contribute to the formation of β-turns due to their exposed hydrophobic proportions is relatively small.

The spike (S) protein of novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is of great concern due to the coronavirus disease 2019 (COVID-19) pandemic. The D614G mutation in SARS-CoV-2 begin to receive widespread attention for its rising dominance worldwide. This mutation changes the amino acid at position 614, from D (aspartic acid) to G (glycine), the initial D614 is now the G614 variant. It is worth noting that the amino acid at position 614 is located at a β-turn in a tertiary structure of the spike. This is consistent with our new theory that both D (aspartic acid) to G (glycine) can result in the β-turn formation. The D-614-G mutation may accelerate the folding of the quaternary structure of the spike due to G614 most likely can contribute to the hydrophobic effect between two tertiary structures of the protein rather than the D614 (see Fig.1D), due to the position 614 located at the docking site in between them.

A typical β-hairpin structure contains two β-strands with hydrophobic attraction between each side-chain and every other side-chain on the strands. Thus, we might be able to predict β-hairpin structures through
evaluating hydrophobic attraction among each side-chain with every other side-chain in the primary structure of a protein. We may can predict β-hairpin through identifying two neighbored sequences of residues in the polypeptide chain both characterized by hydrophobic attraction between each side-chain to every other side-chain, and have two R_B in between them. By using this method, we identified 553 samples in terms of the characteristics above from the 1000 proteins. We find that 158 of the samples are β-hairpins, 36 of the samples are structures of strand-turn-strand, 296 of the samples are structures of strand-turn-helix, 23 of the samples are structures of coil-turn-strand, 23 of the samples are coil-turn-coil and 6 of the samples are α-helices. Thus, physical folding codes for β-hairpins and strand-turn-strand can be deciphered through evaluating hydrophobic interaction among side-chains of an unfolded polypeptide. The results show that strand-turn-helix also can be predict by the method. This indicates that folding of α-helix may be initiated from a β-strand-like thermodynamic metastable state\textsuperscript{15}.

Conclusion

Many amino acid residues contain considerable nonpolar sections in their side-chains, even if they also contain polar or charged groups. This make hydrophobic interaction among neighbored amino acid side-chains in amino acid sequence of polypeptides becomes an important driving force for the stabilization of initial thermodynamic state of unfolded Proteins. The feature of a large hydrophobic surface area covering most side-chains on one side or the other side of adjacent β-strands of a β-sheet is prevail in almost all experimentally determined β-sheets. Minimizing the exposed hydrophobic portions of adjacent side-chains to water should be regarded as the most important driving force for the β-strands formation and caused each side-chains is parallel to every other side-chain on strands. β-turns often contain residues of Aspartate-D, Asparagine-N, Serine-S, Proline-P, Glycine-G which characterized with their side-chains having very small hydrophobic proportions exposure, that is explained as these residues can block hydrophobic effect among neighbored side-chains in sequence, thereby contribute to turns formation. The folding of β-sheets are most likely triggered by multistage hydrophobic interactions among neighbored side-chains of unfolded polypeptides, enable β-sheets fold reproducibly following explicit physical folding codes in aqueous environments. Temperature dependence of the folding of β-sheet is thus attributed to temperature dependence of the strength of the hydrophobicity. The hydrophobic
collapse of β-strands into β-sheets most likely trigger enthalpy-entropy compensation of unfolded polypeptides, enable the main-chain of β-strands to get rid of the hydrogen-bonded water molecules and laterally hydrogen bonding with each other. The folding codes in amino acid sequence that dictate the formation of a β-hairpin can thus be deciphered through evaluating hydrophobic interaction among side-chains of an unfolded polypeptide from a β-strand-like thermodynamic metastable state.

**Materials and Methods**

**Protein structures**

In this study, many experimentally determined native structures of proteins are used to study the folding mechanism of β-sheets. All the three-dimensional (3D) structure data of protein molecules are resourced from the PDB database. IDs of these proteins according to PDB database are marked in the Fig.2, Fig.3, and Fig.4. In order to show the distribution of hydrophobic areas on the surface of β-strands and β-sheets in these figures, we used the structural biology visualization software PyMOL to display the hydrophobic surface areas of these secondary structures.

**Identification of secondary structures of proteins**

Secondary structures of β-strands, β-turns, β-sheets and α-helices were identified in the 1000 proteins by using the STRIDE software. We also used molecular 3D structure display software PyMOL to confirm the identification of secondary structures of proteins.
Figure 1 Hydrophobic portions of amino acid side-chains (hydrophobic portions are highlighted by green)

**Fig. 2. Hydrophobic attraction among neighbored side-chains of β-strands.** (A) A de novo designed protein (PBDID: 5TPJ). (B) The curved β-sheet of 5TPJ. (C) Hydrophobic attraction among adjacent β-strands via hydrophobic surface of side-chains of the β-sheet (hydrophobic surface is highlighted by
using green surface areas). (D) Hydrophobic surface areas on the 6 β-strands of the sheet (green surface areas).

Fig. 3. (A) Hydrophobic surface areas on the β-strands of the protein 1OUR (hydrophobic surface of side-chains is highlighted by using green surface areas), residues located at turns are highlighted in red color in the sequence of the protein. (B) The parts of the protein (residues 1-33 are highlighted in green, residues 34-71 are highlighted in magenta, residues 72-114 are highlighted in red). Hydrophobic surface areas on the β-strands of the sheet (green surface areas).
Fig. 4 Long β-strands (more than 12 residues) characterized with two adjacent R_B residues located at the middle of the β-strands and curved exactly at their two RB residues in the amino acid sequences.
ACKNOWLEDGEMENTS

Lin Yang is indebted to Daniel Wagner from the Weizmann Institute of Science and Liyong Tong from the University of Sydney for their support and guidance. Lin Yang is grateful for his research experience in the Weizmann Institute of Science for inspiration. The authors acknowledge the financial support from the National Natural Science Foundation of China (Grant 21601054), Shenzhen Science and Technology Program (Grant No. KQTD2016112814303055), Science Foundation of the National Key Laboratory of Science and Technology on Advanced Composites in Special Environments, the Fundamental Research Funds for the Central Universities of China, and the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province of China (Grants UNPYSCT-2017126).

Additional information

The authors declare no competing financial interests.

Author Contributions L. Yang, L. Ye and X. H. formulated the study. X. M., L. Yang, C. H. and L. S. conducted the MD simulation. L. Yang, X. M., C. H., L. S., L. L. and J. L. analyzed the PDB data and coded the protein folding codes. L. Yang, X. M., C. H. and L. S collected and analysed the electric charge and rotational resistance data of side-chains. C. H. wrote programs. L. Yang, L. Ye and X. H. wrote the paper, and all authors contributed to revising it. All authors discussed the results and theoretical interpretations.

References

1 Dill, K. A. & MacCallum, J. L. The Protein-Folding Problem, 50 Years On. Science 338, 1042-1046, doi:10.1126/science.1219021 (2012).
2 Lednev, Igor K. Amyloid Fibrils: the Eighth Wonder of the World in Protein Folding and Aggregation. Biophysical Journal 106, 1433-1435, doi:10.1016/j.bpj.2014.02.007 (2014).

*Corresponding author. E-mail address: linyang@hit.edu.cn (Lin Yang) ¹These authors contributed equally to this work.
Alberts B, J. A., Lewis J, Raff M, Roberts K, Walters P. Molecular Biology of the Cell. 4th edn, (Garland Science, 2002).

Grishin, N. V. Fold Change in Evolution of Protein Structures. Journal of Structural Biology 134, 167-185, doi:https://doi.org/10.1006/jsbi.2001.4335 (2001).

van den Berg, B., Wain, R., Dobson, C. M. & Ellis, R. Macromolecular crowding perturbs protein refolding kinetics: implications for folding inside the cell. The EMBO Journal 19, 3870-3875, doi:10.1093/emboj/19.15.3870 (2000).

Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. Science 181, 223-230, doi:10.1126/science.181.4096.223 (1973).

Leopold, P. E., Montal, M. & Onuchic, J. N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. Proceedings of the National Academy of Sciences 89, 8721-8725, doi:10.1073/pnas.89.18.8721 (1992).

Zwanzig, R., Szabo, A. & Bagchi, B. Levinthal’s paradox. Proceedings of the National Academy of Sciences 89, 20, doi:10.1073/pnas.89.1.20 (1992).

Dill, K. A. & Chan, H. S. From Levinthal to pathways to funnels. Nature Structural Biology 4, 10, doi:10.1038/nsb0197-10 (1997).

Walther, K. A. et al. Signatures of hydrophobic collapse in extended proteins captured with force spectroscopy. Proceedings of the National Academy of Sciences of the United States of America 104, 7916-7921, doi:10.1073/pnas.0702179104 %J Proceedings of the National Academy of Sciences (2007).

Chaplin, M. (www1.lsbu.ac.uk/water/water_structure_science.html).

Soares, C. M., Teixeira, V. H. & Baptista, A. M. Protein structure and dynamics in nonaqueous solvents: insights from molecular dynamics simulation studies. Biophys J 84, 1628-1641, doi:10.1016/S0006-3495(03)74972-8 (2003).

Zhang, L. et al. Mapping hydration dynamics around a protein surface. Proceedings of the National Academy of Sciences of the United States of America 104, 18461-18466, doi:10.1073/pnas.0707647104 (2007).

Lin, Y. et al. Universal Initial Thermodynamic Metastable state of Unfolded Proteins. Progress in biochemistry and biophysics 46, 8, doi:10.1647/j.pibb.2019.0111 (2019).

Qiao, B., Jiménez-Ángeles, F., Nguyen, T. D. & Olvera de la Cruz, M. Water follows polar and nonpolar protein surface domains. Proceedings of the National Academy of Sciences 116, 19274-19281, doi:10.1073/pnas.1910225116 %J Proceedings of the National Academy of Sciences (2019).

McPherson, A. & Gavira, J. A. Introduction to protein crystallization. Acta Crystallogr F Struct Biol Comm 70, 2-20, doi:10.1107/S2053230X13033141 (2014).

Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. Nature Structural Biology 10, 980, doi:10.1038/nstb1203-980 (2003).

Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Research 35, D301-D303, doi:10.1093/nar/gkl971 (2007).

Brooks, B. R. et al. CHARMM: the biomolecular simulation program. J Comput Chem 30, 1545-1614, doi:10.1002/jcc.21287 (2009).

Yang, L. et al. Structure relaxation via long trajectories made stable. Phys Chem Chem Phys 19, 24478-24484, doi:10.1039/c7cp04838f (2017).

Uversky, V. N. Intrinsically Disordered Proteins and Their “Mysterious” (Meta)Physics. Frontiers in Physics 7, doi:10.3389/fphy.2019.00010 (2019).
Compiani, M. & Capriotti, E. Computational and Theoretical Methods for Protein Folding. *Biochemistry* 52, 8601-8624, doi:10.1021/bi4001529 (2013).

Callaway, D. J. Solvent-induced organization: a physical model of folding myoglobin. *Proteins* 20, 124-138, doi:10.1002/pro.340200203 (1994).

Rose, G. D., Fleming, P. J., Banavar, J. R. & Maritan, A. A backbone-based theory of protein folding. *Proceedings of the National Academy of Sciences of the United States of America* 103, 16623-16633, doi:10.1073/pnas.0606843103 (2006).

Voet, D. V. J., Pratt. *Fundamental of Biochemistry: Life at the Molecular Level* 4th edn, (Wiley & Sons, 1999).

Pace, C. N., Shirley, B. A., McNutt, M. & Gajiwala, K. Forces contributing to the conformational stability of proteins. 10, 75-83, doi:10.1096/fasebj.10.1.8566551 (1996).

Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The Protein Folding Problem. *Annual review of biophysics* 37, 289-316, doi:10.1146/annurev.biophys.37.092707.153558 (2008).

Voet, D. V. J., Pratt *Principles of Biochemistry*. (Wiley, 2016).

Eisenberg, D. Three-dimensional structure of membrane and surface proteins. *Annual review of biochemistry* 53, 595-623 (1984).

Marcos, E. et al. Principles for designing proteins with cavities formed by curved β sheets. *Science* 355, 201-206, doi:10.1126/science.aah7389 %J Science (2017).

Heinig, M. & Frishman, D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic acids research* 32, W500-W502, doi:10.1093/nar/gkh429 (2004).
Eudes, R., Le Tuan, K., Delettré, J., Mornon, J.-P. & Callebaut, I. A generalized analysis of hydrophobic and loop clusters within globular protein sequences. *BMC Structural Biology* 7, 2, doi:10.1186/1472-6807-7-2 (2007).
### Supplementary Information

#### Note S1. Protein structures samples

| 1B4L | 1KP4 | 1Z3L | 3C3G | 41C9 | 1FDJ | 1R75 | 20A1 | 30ZZ | 5HBP |
| 1B5C | 1KSM | 1Z3M | 3C4S | 41DL | 1FE3 | 1RH9 | 20JR | 3P2X | 5HQQ |
| 1A0B | 1KTH | 1Z3P | 3C5K | 41P1 | 1FER | 1RAQ | 20L1 | 3PAZ | 5HDG |
| 1A18 | 1KX1 | 1Z1A | 3C7I | 41P6 | 1RE1 | 1B11 | 20N8 | 3Q1D | 5HJC |
| 1A6F | 1KXW | 1Z1B | 3C97 | 41PF | 1RES | 1RBW | 20PY | 3Q4Y | 5AKN |
| 1A12 | 1KXX | 1ZJ7 | 3CE8 | 4JHB | 1FEV | 1BXX | 20KQ | 3Q7Y | 5HMB |
| 1AB0 | 1KXY | 1ZPA | 3CEC | 4JJD | 1FF2 | 1RDS | 20SN | 3Q7D | 5HPA |
| 1ACD | 1L5D | 208L | 3CQ1 | 4JZ3 | 1FKK | 1REX | 20UB | 3R3K | 5HQL |
| 1ACF | 1L7L | 2A3G | 3CR2 | 4JZP | 1FKD | 1RFP | 20UM | 3R5P | 5HQL |
| 1AE2 | 1LAC | 2A7B | 3CR5 | 4KV1 | 1FKF | 1RIH | 20X3 | 3RH1 | 5172 |
| 1AF5 | 1LCJ | 2A9I | 3CTG | 4K59 | 1FKK | 1RLK | 20Y3 | 3RHE | 5J5X |
| 1AH0 | 1LEA | 2A9Q | 3CYI | 4KGT | 1FKY | 1RMD | 20Z9 | 3RNT | 51OE |
| 1ALB | 1LGP | 2A1F | 3DQJ | 4KUO | 1FLQ | 1RNN | 20FD | 3RSD | 5KEL |
| 1ANG | 1LKL | 2APB | 3DJU | 4KV4 | 1FLU | 1RQ9 | 20PF | 3RSK | 5KAZ |
| 1LAN | 1LOZ | 2AZ8 | 3DML | 4LBB | 1FLW | 1RNU | 21PL | 3RTO | 5K6M |
| 1AOJ | 1LP1 | 2AZ9 | 3DP5 | 4LFQ | 1FMY | 1RVN | 2P2E | 3RVC | 5KUE |
| 1AQT | 1LRA | 2A3E | 3EAZ | 4LF5 | 1FSN | 1RNV | 2P63 | 3S3Y | 5KXH |
| 1AWJ | 1LSL | 2B1Y | 3ENU | 4LJN | 1FNJ | 1RNX | 2P64 | 3S8S | 5L8Z |
| 1AYC | 1LVE | 2B29 | 3ERS | 4LTT | 1FNK | 1RNZ | 2P6V | 3S9K | 5LAW |
| 1B0T | 1LXI | 2B4A | 3ETW | 4LYO | 1FOW | 1RRI | 2P8V | 3SGP | 5LAL |
| 1B1E | 1LYO | 2B8G | 3EZM | 4MDQ | 1FOY | 1RYR | 2PAL | 3STM | 5L2N |
| 1B1I | 1LZ4 | 2B9D | 3F3Q | 4MJJ | 1G2S | 1RS2 | 2PK7 | 3SUL | 5M9A |
| 1B1J | 1M1S | 2BEZ | 3F45 | 4ML2 | 1GBQ | 1RS1 | 2P9T | 3T1X | 5MXY |
| 1B1U | 1M4A | 2BFH | 3F8C | 4M2Z | 1GD6 | 1RTU | 2PNE | 3T3J | 5NGN |
| 1B2O | 1M4B | 2BHK | 3F9J | 4NOZ | 1GDC | 1RWW | 2PP1 | 3T8R | 5NWX |
| 1B6E | 1M4M | 2BHO | 3FFY | 4NST | 1GHJ | 1RZY | 2PPN | 3T8T | 503A |
| 1BAS | 1MB3 | 2BO1 | 3FRV | 4N6J | 1GKH | 1S3P | 2P2T | 3UA8 | 5C4C |
| 1BEA | 1MG6 | 2BPP | 3FWU | 4NEJ | 1GJM | 1S71 | 2PW5 | 3UB2 | 5C08 |
| 1BEL | 1MH7 | 2BQQ | 3FYR | 4NXR | 1GOD | 1SDZ | 2PWS | 3UB3 | 5MD0 |
| 1BFE | 1MH8 | 2BGF | 3FZ9 | 4OH0 | 1GP3 | 1SF6 | 2P4K | 3UB4 | 5PAL |
| 1BG1 | 1MK0 | 2BS5 | 3FZA | 4OV1 | 1GS3 | 1SF7 | 2PZW | 3UMD | 5PAZ |
| 1BH7 | 1MKU | 2BT1 | 3G7C | 4OXW | 1GSW | 1SF8 | 2QIM | 3UME | 5HRN |
| 1BFK | 1ML8 | 2BWK | 3GBQ | 4OZL | 1GV2 | 1SFQ | 2QAS | 3UNN | 5RNT |
| 1BKV | 1MLI | 2BWL | 3GK2 | 4P15 | 1GVP | 1SK2 | 2QDB | 3V19 | 5TAB |
| 1BM2 | 1N9N | 2BZY | 3GK4 | 4P2P | 1GXT | 1SNP | 2QHE | 3V1G | 5U9U |
| 1BMG | 1N9O | 2CDS | 3GKY | 4P7U | 1H2P | 1SNQ | 2QHW | 3VXX | 5UEP |
| 1BOO | 1NEH | 2C0Q | 3GLW | 4P9E | 1HDO | 1SSC | 2Q1U | 3YVA | 5EUR |
| 1BPP | 1NEQ | 2CW4 | 3GM2 | 4P9V | 1HE7 | 1SV9 | 2QNW | 3W1T | 5UES |
| 1BQK | 1NK0 | 2CXY | 3G3M | 4PAZ | 1HEH | 1T00 | 2QR3 | 3WRP | 5UET |
Table. S1. Randomly selected 1000 small protein structures in PDB

Note S2. Amino acid sequences of β-turns of the 1000 protein samples

| PDBID: 1AA2 | AGYPNV | NFT | RDG | RPDLI | KKS | TKL | VDH | PDBID: 1NXV | PD | LMLPEID | AD | KPF |
|------------|--------|-----|-----|-------|-----|-----|-----|-------------|---|----------|----|-----|
| PDBID: 1ACF | N | GAVT | LDG | SAGF | AG | DDR | GS | EEW | KI | NAENPRG | SETTKGA | NAK | LDSG | SR |
| PDBID: 1A0J | TSK | NEKI | | | | | | | | | | | | |
| PDBID: 1A0J | NSS | MKD | ASG | NNI | TPE | NSSE | MKD | | | | | | | |
| PDBID: 1AYC | DSR | NASG | | | | | | | | | | | | |
| PDBID: 1AYC | RRW | HPNI | VDG | KSNPG | NG | TGDY | LGGG | EEW | KI | NAENPRG | SETTKGA | NAK | LDSG | SR |
| PDBID: 1AYC | IHQQ | KNG | | | | | | | | | | | | |
| PDBID: 1AYC | PDBID: 1A0J | AFSGILA | AADS | DQDK | SAGA | DSDG | | | | | | | | |
| PDBID: 1CKA | DEE | KKG | KPEEQ | DSEG | | | | | | | | | | |
| PDBID: 1FB7 | PDBID: 1FB8 | LWQR | GG | DTGA | IG | CG | PTPYN | | | | | | | |
| PDBID: 1FB7 | PDBID: 1FB8 | SLGT | GGLVK | RN | KIQMS | YSQERV | PF | | | | | | | |
| PDBID: 1FES | CSGC6GA | | | | | | | | | | | | |
PDBID: 1FNK
RD T PD L SGM V TQD K
PDBID: 1FOY
TF T PD S KGR Q
PDBID: 1GBQ
AD D QG D MKQ F
PDBID: 1GVP
KPSQA SRQG NEY DQG QFG DRL
PDBID: 1J2V
DPD KRGD DQN NG KNY
PDBID: 1J41
DGR M QDE GKEV SVG AT PGI
PPH
PDBID: 1J82
NLTDR CKNG TGSS YP GNQY
PDBID: 1K5A
DAKQPQK LTQPCFK ENKN TNP WPC NG QSA
PDBID: 1LEA
FRS VSGAS
PDBID: 1LGP
RLGAEFG KR RRGK DKSG STSG VKQQ QTQ YRNKE
PDBID: 1MH8
VPAR GQ KGR N
PDBID: 1MR0
TLNN KD HSS GNV PYE SKIN
PDBID: 1N0
LYD TET KKG NTEG LTG
PDBID: 1N1
RDD DD AS WNGV TPKD W
PDBID: 10A
EEW KI NAENPRG SEETKG NAK LDSR SR
PDBID: 10P
ADGLCHR
PDBID: 10Y
DPFGQ SHNG NG DEHG
PDBID: 1P9G
CPRP NAG IYD GAGN
PDBID: 1RDS
GS DD DYEG MSTDY GDD HTGASGD
PDBID: 1S3P
AADS GLKKK DKDK SSDA DKDG
PDBID: 1SKZ
PADV DYSGR SGDPKLR RMDH KD DS
DVK
PDBID: 1PKS
LYD REE HLG SDG
PDBID: 1PO8
GCK IDG EP DSKD
PDBID: 1PZ
AE PA NPG VDK IKM PEGA KINE
CPHI GDSP
PDBID: 1PZ
AE PA NPG VDK IKM PEGA KINE
CPHI GDSP
PDBID: 1RNNQ
DSSTS NLTDR CKNG TGSS YP GNPY
PDBID: 1RN
NLTDR CKNG TGSS YP GNPY
PDBID: 1RX
DSSTS NLTDR CKNG TGSS YP GNPY
PDBID: 1RK
YNN ENPP K
PDBID: 1UIG
LDN NF TQA NTDG GLLQ SRWW DGRTPGS
NLCN SSD DGN KGD TGC
PDBID: 1US
PNN TTQAC NSKGG HQLD KSGD TPKG KGL
PDBID: 1WJ
FK PDYI GTRG SLFL
PDBID: 2BEZ
NVLY ISGI NESL
PDBID: 2BPP
IPS N GY KVL NPYTN NN SSEN
PDBID: 2BTI
VL GDE GN PKEV VG GDE GN
PKEV
PDBID: 2CXY
GK PENV
PDBID: 2DN
QGDV LSGN NLDKVS SSSATG IKDY PEDT PAGS
GGGTGLV DTKG NSEYV SGG DSSK DNNG SLGG
QGDV LSGN NLDKVS SSSATG IKDY PEDT PAGS
GGGTGLV DTKG NSEYV SGG DSSK DNNG SLGG
PDBID: 5UEY
AKK DVEALG IKHP PPDH
PDBID: 5USV
TSI FVNQ
PDBID: 5YV7
EDKSPDS G6
PDBID: 6B25
PQENE RPG NED QD ANF QQNE VRT
KEN DG G6
PDBID: 61QC
KIP
PDBID: 9RAT
DSST NLTKDR CKNG YP GNPY
PDBID: 9RNT
CD G6 GS GSNS NYEGLDF LSSG ENN HTGASONNKKYK NDAK PEND NGNE PDTA LNG DFSEVS
PDBID: 1AE2
KPSOA SRQG NEYP DEGQ GQFG DRL
PDBID: 1AIO
DDVN PY PDFV
PDBID: 1B1U
TPSG RL1Q PG LVTEVEC
N
PDBID: 1B1H
TPSW LDN NF TQA NTDG 611Q SRWW DGRTPS
PDBID: 1B3G
SNLNGD NG AKS TQG GED
PDBID: 1ECW
SVLS LRPDD TGTA
PDBID: 1E1G
VSRR PENR RSTC KKG DPKQ KK
PDBID: 1FD8
VM CSQ PD LEKQ
PDBID: 1FDB
IKCK CPV6 PN HPDEC CPAQ EVW DGVKGK
PDBID: 1FKK
PGDGRT KRG EDG SRDRM GKEV G6 SV6 ATG
PDBID: 1FKK
PG1 PPN

SVG GVPDKG DPN GDII DPN NDL
PDBID: 1AWJ
EETL DPQ6 DKNG8E SSYL
PDBID: 1B1E
QDN DAKPQQ6 TSPQD ENKN REN G6S WPPC
NG
PDBID: 1B6A
GPDA PG
PDBID: 1BMG
BHP DG PP NG KS KD6 NSK6
VTLQEP
PDBID: 1CIN
WMPMND DGAMSAL EA KEN PF
PDBID: 1COF
IKCK EVAPVD PN HPDEC CPAQ DGVKG
PDBID: 1D65
DFFQ SHI6 NG DEHG
PDBID: 1E6K
MPNMD DGAMSAL EAK PF
PDBID: 1FD2
IKCK EVAPVD PN HPDEC CPAQ DGVKG
PDBID: 1F56
LDP NF TQA NT6D 611Q SRWW DGRTPS
PDBID: 1H00
NLCN SSD G1LQ KGTD RGC
PDBID: 1HIK
HKCD TLCTE D1FAASK N DTRC PVK
PDBID: 1HRK
PAN AADD QAATKS QF ADAAGA DE QL
VNG CASW
PDBID: 1I99
LDN NF TQA NT6D 611Q SRWW DGRTPS
PDBID: 1JER
NLCN SSD G1N KGTD RGC
PDBID: 1JER
IKCK EVAPVD PN HPDEC CPAQ DGVKG
PDBID: 1K6Q
VG PSSP VG PANA FVNSIND V6TH
PDBID: 1L7L
R6V N6GD GN GSNS R6GR EGTE
PDBID: 6B25
PDBID: 9RNT
KIP
PDBID: 9RAT
DSST NLTKDR CKNG YP GNPY
PDBID: 9RNT
CD G6 GS GSNS NYEGLDF LSSG ENN HTGASONNKKYK NDAK PEND NGNE PDTA LNG DFSEVS
PDBID: 1AE2
KPSOA SRQG NEYP DEGQ GQFG DRL
PDBID: 1AIO
DDVN PY PDFV
PDBID: 1B1U
TPSG RL1Q PG LVTEVEC
N
PDBID: 1B1H
TPSW LDN NF TQA NTDG 611Q SRWW DGRTPS
PDBID: 1B3G
SNLNGD NG AKS TQG GED
PDBID: 1ECW
SVLS LRPDD TGTA
PDBID: 1E1G
VSRR PENR RSTC KKG DPKQ KK
PDBID: 1FD8
VM CSQ PD LEKQ
PDBID: 1FDB
IKCK CPV6 PN HPDEC CPAQ EVW DGVKGK
PDBID: 1FKK
PGDGRT KRG EDG SRDRM GKEV G6 SV6 ATG
PDBID: 1FKK
PG1 PPN

SVG GVPDKG DPN GDII DPN NDL
PDBID: 1AWJ
EETL DPQ6 DKNG8E SSYL
PDBID: 1B1E
QDN DAKPQQ6 TSPQD ENKN REN G6S WPPC
NG
PDBID: 1B6A
GPDA PG
PDBID: 1BMG
BHP DG PP NG KS KD6 NSK6
VTLQEP
PDBID: 1CIN
WMPMND DGAMSAL EA KEN PF
PDBID: 1COF
IKCK EVAPVD PN HPDEC CPAQ DGVKG
PDBID: 1D65
DFFQ SHI6 NG DEHG
PDBID: 1E6K
MPNMD DGAMSAL EAK PF
PDBID: 1FD2
IKCK EVAPVD PN HPDEC CPAQ DGVKG
PDBID: 1F56
LDP NF TQA NT6D 611Q SRWW DGRTPS
PDBID: 1H00
NLCN SSD G1LQ KGTD RGC
PDBID: 1HIK
HKCD TLCTE D1FAASK N DTRC PVK
PDBID: 1HRK
PAN AADD QAATKS QF ADAAGA DE QL
VNG CASW
PDBID: 1I99
LDN NF TQA NT6D 611Q SRWW DGRTPS
PDBID: 1JER
NLCN SSD G1N KGTD RGC
PDBID: 1JER
IKCK EVAPVD PN HPDEC CPAQ DGVKG
PDBID: 1K6Q
VG PSSP VG PANA FVNSIND V6TH
PDBID: 1L7L
PDBID: 1FNJ
RDT TPDL LSGWQYV VTGGLKK PQDQI
PDBID: 1GDC
CLVCS YG CAGRN NLEA
PDBID: 1IZ
AR TDK NENQ GLFQ DRYW GST GKDCN
TDD KFDIW NIS
PDBID: 1KL
CQCIKT HPKF GPHCA SDG DPKE
PDBID: 1LRQ
PDK
PDBID: 1LYU
KTG EVEQ SAKA SPKA KLGD EG PAAGA
PDBID: 1J81
NLTKDR CKNG TGSS YP GNPY
PDBID: 1K40
PAS
PDBID: 1KL
EPW KNL APMTHHG SESTAGS QNQ LDNG PR
SDGLCTR
PDBID: 1LRA
CD GS GSNS NYEG LSSG ENN HTGASGNNAE PA NPG VDK IKDM PEGA KINE
PDBID: 1N90
ATLPDC GPDEVLG GEGT RKDG TPDG
PDBID: 1NEQ
NEKARD LER IWPSRYQ
PDBID: 1NK0
EGM DSD NPAW RDR RDA GN
PDBID: 1O4M
EEW KI NENFRG SETTKGA NAK LDSG SR
ADGLCHR
PDBID: 107Z
CIS NPRS SQFCP KKKGE NPES CISI SQFCP
KGE NPES
PDBID: 1P6S
SDSG SDSG
PDBID: 1PZ4
GIRM PA GG LKNV AE GALP
PDBID: 1QLS
RDGNNT TE DLDS
PDBID: 1R9H
TPKK GG TTG ENG GRGNVI DA PK
PDBID: 1UHD
ANNEAG NPGD YGP REH NS VNT PNNV
PDBID: 1LRQ
VPGTYGN
PDBID: 1LZ4
MDG GY TRA AGDR G1FQ SRY DGKTPGA
AHH QDN DPQ QNR YQGC
PDBID: 144B
NYKNPK KKA
PDBID: 1M8
QIWF FKNW CACT TENEPI GDFF EPDD SPGC
VKK
PDBID: 1M8
MQA GL ASG GNSDKA QD RDL
KAV
PDBID: 1N90
ATLPDC GEGT RKDG TPDG
PDBID: 1001
FNF NKKG AD6Q
PDBID: 1018
RLG NTDO QTG RWNI KWYT KDHV LGDYE
HDG
PDBID: 1PAZ
CTPH GDSP
PDBID: 1PBI
CDT SNPP SNPP QCIN KNKSA CDT SNPP
SNPP QCIN KN
PDBID: 1QPV
KYYK NDAK NGNE PDTA LNGV DFSEVS
PDBID: 1RNZ
DSSTS NLTKDR CKNG TGSS YP GNPY
PDBID: 1RRY
NVIDT YNR ENP
PDBID: 1RSI
NVIDT YNR ENPP1
PDBID: 1SFB
LDN RG NF TQA NTDG YLG1LQ NSRWWC
DGRTPGS NLCS NLSSDI VSDG CKGT W1R
PDBID: 1U3Z
DRT VTG KD EENG RQF YKDVNI RKS
PDBID: 1U42
DRT CVTG QKDD1 EGV SPTDV YKDVNI KSD
PDBID: 1UHD
| PDB ID | Protein Description |
|--------|---------------------|
| 4DP4   | GADDGSLA            |
| 4EES   | ADDG                |
| 4ETD   | SHG PAF PPNKQ PLS LNGE LRNG |
| 4GBN   | PDBID:4GBN          |
| 4EC2   | PDBID:4EC2          |
| 4ETD   | TSI PBDID:4ETD      |
| 4IDL   | QTTG APGK TTG GDFVGR NANN DSL DGAR |
| 4J23   | DPN LYD ATLS TKG NHNGE KN  |
| 4KV4   | DVEALGLH IKHP PPDH |
| 4MDQ   | DEKQGH CSND VKE |
| 4QYW   | PDBID:4QYW          |
| 4X02   | PD MPEMN DPNA MG PF |
| 4C4P   | PDBID:4C4P          |
| 4DP    | PD PG AEPP GAGE NQD GPGG |
| 5DKN   | EGDKH SHFL DSDG FFE |
| 5EI4   | PDBID:5EI4          |
| 5HDG   | EPE PBDID:5HDG      |
| 5SWX   | DAL SPSG FD PC RG  |
| 5OC8   | PDBID:5OC8          |
| 5UEV   | EK CSND VKE         |
| 5UVY   | PDBID:5UVY          |
| 5CPV   | DVEALGLH IKHP PPDH |
| 6EVL   | PDBID:6EVL          |
| 6PAZ   | DPSPH PBDID:6PAZ    |
| PDBID: 1GD6 | PDBID: 1EQV | PDBID: 1FIK |
|-------------|-------------|-------------|
| SR TSK NRGN GLFQ DRYW KDCN TDD | GD KG LL ETK PKK AKK DKG | DG AP |
| RFDAW HCQGS | | PDBID: 1FKD |
| PDBID: 1GJH | | YKDSP VPGK GG LLQDGEF TGGA TDK GV |
| PESIA WHKKPE KRD DKV NEGD LSG | | DGRRT KRG EDG GK SVG ATG PGI |
| PDBID: 1GHJ | | PPH |
| AQ SRQG NEYP DEGQ GQFG DRL | | PDBID: 1FKF |
| GPKR LKGR GNR LPGR | | DGRRT KRG EDG GK SVG ATG PGI |
| PDBID: 1GHT | | PPH |
| VQGV DGDG PPLA SQL | | PDBID: 1HEY |
| PDBID: 1H2P | | |
| RN VPQG FG NTG ST ISGSS IDN | | DKEL TGKS NN WNMPMD DGAMSAL KPF |
| QGERDH YR NKGF EH NNDE | | PDBID: 1HME |
| PDBID: 1ID1 | | DPNA HPGL |
| TAT NLCY CDAF SRGK KPYE TDKC | | PDBID: 1IJO |
| PDBID: 1I7Z | | |
| NLTQD CRNG TGSS XP GPNI | | QDGN TGVI PRPC NQ GR |
| PDBID: 1JPO | | PDBID: 1KMM |
| LDN NF TQA NTGD GILQ SRWW DGRTPGS | | QH EFY DAG |
| NLCN SSD KGTD RGC | | PDBID: 1KMS |
| PDBID: 1KM8 | | |
| DN GG TGV PRPC NQ GR | | DN GG TGVI PRPC NQ GR |
| PDBID: 1KSM | | PDBID: 1KXX |
| KEGDPNQL DKNG | | LNN NF TQA NTGD GILQ SRWW DGRTPGS |
| PDBID: 1KTH | | NLCN SSD DGN KGTD RGC |
| PDBID: 1LJO | | PDBID: 1LXI |
| PNTK NENK | | WDQW AFPL NPETV DDSS |
| PDBID: 1MG6 | | PDBID: 1MG6 |
| MDG GY TRA AGDR GTFQ SRYW DGKTPGA | | GCNC NK IL |
| NACH QDN DPQ QNR VQGC | | PDBID: 1MME |
| PDBID: 1Q2E | | |
| VPAR GCY GG QG KGRN NI | | IPSS NN GCY NPYTN NN SSEN NLDMKN |
| PDBID: 1Q4K | | PDBID: 1Q45 |
| EEW KI NAENPRG SETTKGA NAK LDSG SR | | EEW KI NAENPRG SETTKGA NAK LDSG SR |
| ADGLCHR | | ADGLCHR |
| PDBID: 1Q4L | | PDBID: 1Q8B |
| EEW KI NAENPRG SETTKGA NAK LDSG SR | | ADGLCHR |
| ADGLCHR | | PDBID: 1Q8B |
| PDBID: 1Q8D | | DLNEI EEE |
| EEW KI NAENPRG SETTKGA NAK LDSG SR | | PDBID: 1R26 |
| ADGLCHR | | MRAR AVWC FPTV ADNN QLP SG GAN |
| PDBID: 1R6U | | PDBID: 1RLK |
| PQ GG YP SEDI VYNG RD TNTG | | KDLD GYQVE |
DG EGI
PDBID:2V6H
TVGG GKW QH RASK TDA TKDK
PDBID:3A0V
KDG VLG LPD KGER NAKTQ
PDBID:3B19
LGQ PNSKCN A DGTR STK KVQFG SNT GWFN
AL LV
PDBID:3ENU
KNF PRL PFG WENK GPR HNY DAGA
ANL FDNF
PDBID:3EZM
GS RTNG LNSV DG NFIET GSS TRAQ
DG
PDBID:3FAJ
KKGD
PDBID:3GLW
RNF TR GQ
PDBID:3GM3
RTDD AQL AVT
PDBID:3GU
PSS LGAR PGW IAE
PDBID:313Z
TSI FVNQ
PDBID:317W
DSSTS NLTKDR KP CKNG TGSS YP GNPY
PDBID:317X
DSSTS NLTKDR KP CKNG TGSS YP GNPY
PDBID:3IE9
ADGA KM ETP KVG VAGVLG KKE TP
PDBID:31NC
TSI
PDBID:3I4G
LDNYRGY NF TQA NTDGS GILQ SRWW TPGS
PDBID:3J2R
NLCN SSD KGT IRGC
PDBID:3JZR
KDT DEKQKH CSND VKE
PDBID:3KVT
GG KIPAT TEGMLN PVLN PTDVC
PDBID:3ONH
PQD ASNQ YD EDLNDR GNG EEGDT DELPCNT
PDBID:3S9K
NLETYW SI KEG SRTPGT KAI1SEN DSPK FDS
PDBID:3PAZ
PDBID:2OQK
EEG NG FDG NPG DF GEIPETT
PDBID:2OUB
SS GCY GWG NG LY
PDBID:2P61
SPT KL EWQTI
PDBID:2PNE
CPG GPG NPCC GT GTPK
PDBID:2PVTT
GCY GG NG LY
PDBID:2Q0B
GN KGQ LL AKK DKYG DG KG
PDBID:2Q0W
GCY WG NG LY
PDBID:2REA
DGR YNPD EIQ RMVGR
PDBID:2REY
MEYN MKGG EPG ND FENM
PDBID:2UX
P1GW MGNC
PDBID:2VSL
PRN FGTV GD FHCG KPSE PGC
PDBID:2W51
RPGDC TFS TDDA DSQCEL KKL GETCKGC
PDBID:3BKS
BRSKR EG FNDL ELAV TADYP KETS
PDBID:3CQ1
DPELG VNVLG PP PLHD LPGV FEPP RLL
PDBID:3DML
QFGC QRD PPGL LARP FTP GD
PDBID:3GC7
WIREYP REES
PDBID:3KL
LSFFPGQ7 PT1KRF DKEG TTEL PTFKA
PDBID:3L1M
GK
PDBID:3LLH
E PN GD GQP GD PS
PDBID:3P2X
LYFQGL FAGR NECH KS VPEV
PDBID:3PAZ
PDB ID: 1FLQ
LDN NF TQA NTDG G1LQ SRWW DGRTPGSDMLK
NLCN SSD DGGHQNAWKTAD RGC

PDB ID: 1FLY
LDN NF TQA NTDG G1LQ SRWW DGRTPGSDMLK
NLCN SSD DAN KGTG RGC

PDB ID: 1GOD
KDKT

PDB ID: 1HEH

PDB ID: 1HRC

PDB ID: 1I73

PDB ID: 1J7A

PDB ID: 1K1Z

PDB ID: 1MG

PDB ID: 1M1S

PDB ID: 1O41

PDB ID: 1PAF

PDB ID: 1PI4

ADGLCHR

PDB ID: 1I0F

ADGLCHR

PDB ID: 1O65

SALT EDDSV RYG

PDB ID: 1PAL

SFAG AADS DQDK FSPSA DRRG

PDB ID: 1Q4R

KDGV H1QG

PDB ID: 1QJL2

IDT IDT IDT

PDB ID: 1T00

IMAG TDD LKN AAAC GDKI IDEN SIP

PDB ID: 1UIA

SKHAFSL CSKC REN LTDG YFDG PD YEDD

PDB ID: 6INS

PDB ID: 6JQ4

PDB ID: 1A0B

PDB ID: 1AF5

PDB ID: 1B11

PDB ID: 1BAS

PDB ID: 1CYI

PDB ID: 1DKJ

PDB ID: 1EV3

PDB ID: 1H7E

PDB ID: 1IQ8

PDB ID: 1M8

PDB ID: 1N0V

PDB ID: 1OR5

PDB ID: 1OR

PDB ID: 1P00

PDB ID: 1R0W

PDB ID: 1T00

PDB ID: 1U1A

SKHAFSL CSKC REN LTDG YFDG PD YEDD

PDB ID: 6INS

PDB ID: 6JQ4

PDB ID: 1A0B

PDB ID: 1AF5

PDB ID: 1B11

PDB ID: 1BAS

PDB ID: 1CYI

PDB ID: 1DKJ

PDB ID: 1EV3

PDB ID: 1H7E

PDB ID: 1IQ8

PDB ID: 1M8

PDB ID: 1N0V

PDB ID: 1OR5

PDB ID: 1OR

PDB ID: 1P00

PDB ID: 1R0W

PDB ID: 1T00

PDB ID: 1U1A
| PDBID: 1UWM | PDBID: 1KH0 |
|-------------|-------------|
| DN NF TQA NTDG G1LQ SRWW DGRTPGS | MDG GY TRA AGDR GIFQ SRYW DGKTPGA |
| NLCN SSD GN KGT RGC | NACH QDN DPQ QNR VQGC |
| EHNG KPG AYEPNATDG | FANG NG FANG NG |
| WDDW PL EFPL DPEST DSAN | DSDTS NLTKDR CKNG TGSS YP |
| PDBID: 1WQ | PDBID: 1KH8 |
| PDBID: 1WHI | DSSTS NLTKDR CKNG TGSS YP |
| PDBID: 1WY9 | | |
| GD | | |
| SNC NGNG SEE | | |
| PDBID: 1YGT | | |
| GGN KP NNDTD NKT | | |
| PDBID: 1Z21 | | |
| SNPP | | |
| PDBID: 2A9J | | |
| TPST PQE KPSG RYN LQTGL | | |
| PDBID: 2AZ8 | | |
| LWKR P GGG GTAG LPG IG CG GPT | | |
| PDBID: 2BQQ | | |
| NG RYF SSDRF N1 LPQG IDG EEGE | | |
| PDBID: 2COQ | | |
| DDV NPNW | | |
| PR ETGE DTA LGS CR RDL AFN | | |
| PDBID: 2EH9 | | |
| ELG GGE GEEP | | |
| PDBID: 2HNV | | |
| GD AY RG IG FGD ED PS | | |
| PDBID: 2J1N | | |
| PDBID: 2COQ | | |
| DNLV PS LG KENG QEG NG LKNL GY | | |
| PDBID: 200P | | |
| RGQAN AEPGED DADG | | |
| PDBID: 200Q | | |
| RGQAN GED DADG | | |
| PDBID: 20AI | | |
| REDG GTDT NNYH LAG HVG AG GA | | |
| PDBID: 20J1R | | |
| DTNN EGDE DTNN EGDELLA TLTG EPSD AG | | |
| PDBID: 20M8 | | |
| EDGR LSDYN QKES LR RG | | |
| PDBID: 208 | | |
| GKL | | |
| PDBID: 2P1X | | |
PDBID: 3V19
HSI

PDBID: 3VYA
ED DGN GGM RDE RKV RN IARFKWA

PDBID: 4AHI
DAPKQR GLTSPCID ENKNG REN GGS WPPC NG

PDBID: 4AQI
GRDG DKNE

PDBID: 4AQJ
RRDG NYLA DKNE

PDBID: 4EXO
TSI

PDBID: 4LYO
LDN NF TQA NTDG GILQ SRWW DGRTPGS

PDBID: 4UNG
NLCN SSD GNMGNAW KGTD RGC

PDBID: 5AFG
DAAQHQ CSND VKE

PDBID: 5B1G
LDN NF TQA NTDG GILQ SRWW DGRTPGS

PDBID: 5BMH
LCN SSD GNMGNAW KGT RGC

PDBID: 5C6X
GKTL

PDBID: 5C8X
PD PG AEPP GAGE GPNG

PDBID: 5CUL
SS PTH RRGETPLP NVD

PDBID: 5D53
PS

PDBID: 5FD1
IKCK CPVD PN HPDEC CPAQ DGVKGK

PDBID: 5KAZ
LPY HGRL VDG SESIPG KN EKIKY AEGS

PDBID: 5N5N
DSCEYC CCP DSCSEYC DGQ CCP

PDBID: 5OC4
PAQI AE

PDBID: 5PAZ
AE PA NPG VDK IKDM PEGA KINE

PDBID: 3C5K
PG CPH PAAG DVTQ CGDCG IQE LSCY GRYING

PDBID: 3C7I
YIDL YIQG GEDM

PDBID: 3DJN
HPW KI HDG ESAPG GN DGAG LWV

PDBID: 3FFY
RST SROQ

PDBID: 3GKY
HSI HSI IGERG

PDBID: 3HQB
YKH NNDE NNDE

PDBID: 3IN2
AECS DQM NTN DKSCQ PKNMIG TAAD GSG

PDBID: 3LRO
FPGHSSL ALL

PDBID: 3LYE
DPD EKTD GDDT ND DD

PDBID: 3LYE
PD SLG SQS SSNS KPGQ WAS ESGVPDR

PDBID: 3MF8
QDRLT PAGN GG VEYG

PDBID: 3MYA
PRGVPSR LVNT PRGV PS

PDBID: 3QTY
TD TG NPDG DRSDPGI DNG TDTG NPDG DRSDPGI

PDBID: 3RHE
DGNG TDTG NPDG DRSDPGI

PDBID: 3RHE
KN PIES PT VGTK IEPKA SNE QDF

PDBID: 3STM
PG DE

PDBID: 3SM
KGKD I G K GS TMTG EGDN KN GD

PDBID: 3SU
NADQ VACS GPNG SGTM AGNG NGQ

PDBID: 3T1X
DRKG ALWA PPP LLG GERM GEHA DETA
PDBID: 5TAB
GPLGSEV RCTICE NDF CEEQ CYVC

PDBID: 5UEET
DVEALGL IKHP PPDH

PDBID: 5XUK
KRFK KDK GKELS SPKN AG

PDBID: 5ZND
RDEVA PDCDDW DPHILCD

PDBID: 6CEE
CPH PAAG DYTQ CGDCG IQE LSCY GRYING

PDBID: 6EKB
YIDL YYCQ

PDBID: 6I3S
CANCEEG CSQCKGG HFNGL KAG CWLCRGK CGDCNGA

PDBID: 6LQ9
EKQH CSND VKE

PDBID: 6MQ6
DPRLPDM ILG GPET TKSG DQKG

PDBID: 6RNT
CD GS GSNS NYEGDF LSSG ENN HTGASQNNGR

PDBID: 6B20
TVC NPGT PDDW CPLCA TVC NPGT PDDW

PDBID: 1BKF
DGRT KRG EDG NK GK SYG ATG

PDBID: 1BPQ
PG I PPPH

PDBID: 1B2O
TV C NPGT PDDW CPLCA TVC NPGT PDDW

PDBID: 1COB
KET DSST NLTKDRCKN QKQTN KYPN GPY

PDBID: 1C9H
DGRT KKG QNG SRDRN GK SLG ATG

PDBID: 1DMM
PGV PPN

PDBID: 1DMQ
GD DPFGQ SHING NG DEHG

PDBID: 1DZ0
AQ C NDAM NVK DKSCK AKVAMG GGG FPGHWMMP RDBID: 5CB9

PDBID: 3WW5
LDN NF TQA ETDG GILQ SRW DGRTPGS

PDBID: 3ZEK
LDN NF TQA NTDG GILQ SRWW DGRTPGS

PDBID: 4AHG
TSPC ENKN REN GG WPCC NG

PDBID: 4ET9
LCN SSD GNGMNA W KGD RGC

PDBID: 4FY
SS GCY GWG VNGA LYPDFLCK

PDBID: 4HMB
GCY WG NG LY

PDBID: 4H69
FDQSR ARVENC MQ CTC GPR T RGD NEDG

PDBID: 4HRS
VYGE DDGD DG PEDQ VEY RLIK

PDBID: 4HSW
LDN NF TQA NTDG GILQ SRW DGRTPGS

PDBID: 4KU0
LCN SSD GNGMNA W KGD RGC

PDBID: 4LR6
SCI RKTCG

PDBID: 4ML2
PD TDK

PDBID: 4PSV
TLPL KG PD TDK

PDBID: 4PTS
HPW KI IDG ESAPG GN DGAG LW

PDBID: 4PTA
NW DTGS EKPRN

PDBID: 4RXA
TSI

PDBID: 5AEF
GV V

PDBID: 5CB9
PD PA PG AEPP GAGE NDTACCY GPGG
CTAH  GDSP
PDBID: 4TS8

QDPES  DPQLLG  VKNP  GQYQE  RKTS
PDBID: 4XL

GVOK  KDGP  RGG  FPDG  INGK
PDBID: 5B52

RAL
PDBID: 5ER4

EGDKH  SHFL  DSDG
PDBID: 5GSP

CD  GS  GSNS  NYEG  LSSEG  ENN  HTGASGN  PDBID: 2XDY

PDBID: 5LAZ

EKQII  CSND  VKE
PDBID: 5UES

DVEAGLHI  IKHP  PPDH
PDBID: 5Y11

ATNDERV
PDBID: 6C6D

CPH  PAAG  DVTQ  CGDCG  IQE  LSCY  GRYING
PDBID: 6FLG

Y1DL  YYYQ
PDBID: 6FLG

HDAWPFNL  NPRLVSG  IKNP  EDDS
PDBID: 6H0K

LDN  NF  TQA  NTDG  GILQ  SRWW  DGRTPGS  DQSN  QDQR
PDBID: 155C

NLCN  SSD  DNGMNAYW  KGD  IRGC
PDBID: 1ADY

NEG  KCCAC  GKT  NPDL  GRN  AXXX
PDBID: 1B2F

PDBID: 1A9G

QDN  DAKPQQR  GLTSPCKD  ENKN  REN  WPPC  NG
PDBID: 1FZA

PDBID: 1AQ7

AEQ  SEG  YPGH  QHG  PG  SHG
PDBID: 1BXY

P1GY  RLQ  P1GY  RLQ  AIL
PDBID: 1CDP

FAGVL  AADS  GLTSK  DQOK  FKADA  DSDG
PDBID: 1CDT

KLIP1A  PEKGN  ASKRM  SAL  TDRC  KLIP1A  PEGK
PDBID: 1DPY

ASKRM  NVC  SAL  TDRC
PDBID: 1DPY

G1PY  PG  NPNIK  QP  DSAD  ML  STSC
PDBID: 1DJ7

GDN  PY  PTPVP  DGDN  PY  PTPY
PDBID: 3NGR

CG  KENP  YKNS  PT  TC
PDBID: 2RKN

NLTKDR  CKNG  TGSS  YP  GNPY
PDBID: 2RNS

AV  EA  DV  DPGD  TD  SGA
PDBID: 2VJW

SSSD  G1CT  NDDQ  EASC  EAS  CDTCM  TE
PDBID: 2V7P

GKANEG  HQ  PRK  TFGV  STKG
PDBID: 2XFE

VNG  GG  TG  NG  TGV  NN  GG
PDBID: 2XFS

HCN
PDBID: 2XMU

VPTI  DAQA  LTSK  VPTI  EDAQA  LTSK
PDBID: 2XHY

QVET1VS
PDBID: 3A0D

SPN  TG  GPS  QGDC  SG  TGGLGSG  CG  HNNG
PDBID: 3A0D
| PDB ID | Amino acid sequences |
|--------|----------------------|
| 1FIW   | EEW KI NPENPRG SETTKGA NAK LDSG SR PDBID:3R5P QSGP KN ATN QPGT DVGRNV SVES |
| 1FLU   | ADGLCHR |
| 1FLW   | LDN NF TQA NTDG G1LQ SRWW DARTPGS PDBID:3R8H |
| 1HQB   | SQFG PVSEF WDT PDBID:3RSK |
| 1107   | NLCN SSD DGNMNAWGKTD RGC PDBID:1I07 |
| 1117   | NSS MKD ASG NNI ARNSeE MKD NASG LCN SSD DGNMNAWGKTD RGC PDBID:1HR7 |
| 1118   | NNI |
| 1119   | PDBID:4ET8 |
| 1120   | LDN NF TQA NTDG G1LQ SRWW DGRTPGS DNGMNAWGKTD RGC PDBID:1FLW |
| 1121   | AQNEDEL IKS EGG YGGK PDBID:1117 |
| 1122   | PDBID:4FC1 |
| 1123   | PDYAL |
| 1124   | NLCN SSD DGN KGTG RGC PDBID:4FDX |
| 1125   | NG AQ GASPTA KV PGDNQ AFGG CDAF GVEHF AG GVEHF AG PDBID:4IP6 |
| 1126   | NLCN SSD DGN KGTG RGC PDBID:1118 |
| 1127   | NLCN SSD DGN KGTG RGC PDBID:1120 |
| 1128   | PDBID:4N0Z |
| 1129   | PDBID:1L1P1 |
| 1130   | LDN NF TQA NTDG G1LQ SRWW DGRTPGS DTGNSFS IKAPK AKDT PDBID:1L44A |
| 1131   | PDBID:3HBO |
| 1132   | IAGH HPSS IAGH HPSS IAGH HPSS IAGH PDBID:6BSY |
| 1133   | PDBID:615A |
| 1134   | IAGH HPSS IAGH HPSS IAGH HPSS PDBID:1NN7 |
| 1135   | SG YPDTLGSG PRHE LIPE |

Table. S2. Amino acid sequences of β-turns of the 1000 protein samples.
Note S3. Amino acid sequences of long β-strands with two adjacent RB residues locating at the middle segments of the β-strands.

Figure S1 Long β-strands haven’t curved at their two RB residues in the amino acid sequences in the 1000 proteins.