Comprehensive Folding Variations for Protein Folding

Jiaan Yang[1], Wen Xiang[1], Xiao Fei Zhao[2], Gang Wu[3], Shi Tong[4], Qiyue Hu[5], Hu Ge[5], Qianshan Qin[5], Xinshen Jin[5], Lianshan Zhang[5], and Peng Zhang[1]

[1]Chinese Academy of Sciences Shenzhen Institutes of Advanced Technology
[2]Micro Biotech Ltd Shanghai 200123 China
[3]Huazhong University of Science and Technology Tongji Medical College
[4]Shenzhen Hua Ying Kang Gene Technology Co Ltd Shenzhen Guangdong 518057 China
[5]Shanghai Hengrui Pharmaceutical CoLtd Shanghai 200245 China

February 16, 2022

Abstract

The revelation of protein folding is a challenging subject in both discovery and description. Except acquirement of accurate 3D structure for protein stable state, another big hurdle is how to discover structural flexibility for protein innate character. Even if a huge number of flexible conformations are known, difficulty is how to describe these conformations. A novel approach, protein structure fingerprint, has been developed to expose the comprehensive local folding variations, and then construct folding conformations for entire protein. The backbone of 5 amino acid residues was identified as a universal folden, and then a set of Protein Folding Shape Code (PFSC) was derived for completely covering folding space in alphabetic description. Sequentially, a database was created to collect all possible folding shapes of local folding variations for all permutation of 5 amino acids. Successively, Protein Folding Variation Matrix (PFVM) assembled all possible local folding variations along sequence for a protein, which possesses several prominent features. First, it showed the fluctuation with certain folding patterns along sequence which revealed how the protein folding was related the order of amino acids in sequence. Second, all folding variations for an entire protein can be simultaneously apprehended at a glance within PFVM. Third, all conformations can be determined by local folding variations from PFVM, so total number of conformations is no longer ambiguous for any protein. Finally, the most possible folding conformation and its 3D structure can be acquired according PFVM for protein structure prediction. Therefore, the protein structure fingerprint approach provides a significant means for investigation of protein folding problem.

Comprehensive Folding Variations for Protein Folding

Jiaan Yang[*, #, 1,2], Wen Xiang Cheng[#, 1], Xiao Fei Zhao[2], Gang Wu[3], Shi Tong Sheng[4],

Qiyue Hu[5], Hu Ge[5], Qianshan Qin[5], Xinshen Jin[5], Lianshan Zhang[5], Peng Zhang[1]

[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong, 518055, China

[2] Micro Biotech, Ltd., Shanghai, 200123, China

[3] School of Basic Medicine, Tongji Medical College, Huazhong University of Science and Technology, 430030 Wuhan, China

[4] Shenzhen Hua Ying Kang Gene Technology Co., Ltd, Shenzhen, Guangdong, 518057, China

[5] Shanghai Hengrui Pharmaceutical Co.Ltd., Shanghai 200245, China

[#] Contributed equally to this work

* Corresponding author: jiaanyang@yahoo.com

Keywords: protein folding, protein conformation, folding conformation, protein structure prediction, intrinsically disordered protein, alphabetic description

ABSTRACT

The revelation of protein folding is a challenging subject in both discovery and description. Except acquirement of accurate 3D structure for protein stable state, another big hurdle is how to discover structural flexibility for protein innate character. Even if a huge number of flexible conformations are known, difficulty is how to describe these conformations. A novel approach, protein structure fingerprint, has been developed to expose the comprehensive local folding variations, and then construct folding conformations for entire protein. The backbone of 5 amino acid residues was identified as a universal folden, and then a set of Protein Folding Shape Code (PFSC) was derived for completely covering folding space in alphabetic description. Sequentially, a database was created to collect all possible folding shapes of local folding variations for all permutation of 5 amino acids. Successively, Protein Folding Variation Matrix (PFVM) assembled all possible local folding variations along sequence for a protein, which possesses several prominent features. First, it showed the fluctuation with certain folding patterns along sequence which revealed how the protein folding was related the order of amino acids in sequence. Second, all folding variations for an entire protein can be simultaneously apprehended at a glance within PFVM. Third, all conformations can be determined by local folding variations from PFVM, so total number of conformations is no longer ambiguous for any protein. Finally, the most possible folding conformation and its 3D structure can be acquired according PFVM for protein structure prediction. Therefore, the protein structure fingerprint approach provides a significant means for investigation of protein folding problem.

INTRODUCTION

Protein folding is one of the challenging subjects in science,11Science 1 July:Vol. 309 no. 5731 pp. 78-102 (2005).,22Dill K A, Maccallum J L. The protein-folding problem, 50 years on.[J]. Science, 338(6110):1042 (2012). which particularly has attracted much attention since recent progress by AlphaFold. With artificial intelligence (AI) approach, AlphaFold made a significant breakthrough to accurately predict 3D structures based on protein sequence.33 J. Jumper, etc. "Highly accurate protein structure prediction with AlphaFold". Nature. 596 (7873): 583–589 (2021). ,44Callaway, Ewen. "'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures". Nature. 588 (7837): 203–204 (2020). However, the protein folding problem has not been thoroughly resolved yet because protein is not a static structure.55 Stephen Curry, No, DeepMind has not solved protein folding, Reciprocal Space (blog), 2 December (2020).,66Balls, Phillip. "Behind the screens of AlphaFold". Chemistry World, (2020). The intrinsically disordered protein (IDP) has already discovered that many proteins lacked a fixed three-dimensional structure, and many protein functions were accomplished with ensemble of flexible conformations.77Robin van der Lee and etc, Classification of intrinsically disordered regions and proteins.[J]. Chemical Reviews, 2014, 114(13):6589.,88Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z (2001). "Intrinsically disordered protein". Journal of Molecular Graphics & Modelling. 19 (1): 26–59.,99Dyson HJ, Wright PE (March 2005). "Intrinsically unstructured proteins and their functions". Nature Reviews Molecular Cell Biology. 6 (3): 197–208,1010Dunker AK, Silman I, Uversky VN, Sussman JL (December 2008). "Function and structure of inherently disordered proteins". Current Opinion in Structural Biology. 18 (6): 756–64. Thus, except for accuracy in structure, biologists also want to know how many different ways the protein will fold into, why protein are in such folds and what biologic functions are impacted by folding. To date, it is well known that the protein folding patterns are primarily decided by global multiple weak interactions of protein itself, such as hydrogen bond, disulfide bond, van der Waals force, electrostatic interactions and hydrophobic interactions, etc. Also, it is influenced by environment factors, such as protein-protein interactions, ligands, ions, solvent,pH,temperature and chaperones, etc. Under constraints, a protein can still fold into various conformations between random coil and native state, or undergo reversible folding process between disorder and order transitions. In 1957, Francis Crick indicated

that protein folding was simply a function of the order of amino acids.1111Cobb M. 60 years ago, Francis Crick changed the logic of biology[J]. Plos Biology, 15(9):e2003243, (2017). It is true that different order of amino acids or replacement of residue in sequence may cause the change in folding conformation. In 1969, Cyrus Levinthal indicated that protein may have an astronomical number of local minima in conformational space,1212Levinthal, C. How to Fold Graciously. In Mossbauer Spectroscopy in Biological Systems, pp 22-24, Allerton House, Monticello, IL (1969). and further pointed out to understand the relationship from sequence to protein folding was a challenging problem. The basic task includes how to obtain all possible folding conformations, how to present these folding conformations with an astronomical number and how to acquire the most possible conformation in stable state and its 3D structure.

In spite of the lack of systematical approach, some protein conformations are still known. The protein conformations may be obtained by protein 3D structures which are experimental measurement data or results of computational approaches. Experimental measurements, such as X-ray crystallography, Nuclear Magnetic Resonance (NMR) and Transmission Electron Cryomicroscopy (CryoTEM) etc., may accurately determine atomic coordinates of protein 3D structures. However, they only provided the limited folding conformations for protein stable states under specific conditions. Anyhow, the results from experiments are snapshots of protein structures which provide significant folding information, but they couldn't cover the enormous conformational space. Also, the progress of experimental measurements for protein 3D structures cannot keep up with the pace of rapid increase of knowledge of protein sequences as a huge number of protein sequences are determined by genetic code. To date, over 35,000,000 gene codes are available in National Center for Biotechnology Information (NCBI) database,1313https://www.ncbi.nlm.nih.gov/ and over 225,000,000 protein sequences in Universal Protein Resource (UniProt) database.1414http://www.uniprot.org/ So far, merely about 187,000 of 3D structures are available in Protein Data Bank (PDB).1515https://www.rcsb.org/ In other words, less than 1% of total protein sequences have the known protein 3D structures. Therefore, on the other hand, the development of computational approaches becomes an important methodology to predict the protein 3D structures. The effort of protein structure prediction, however, is primarily focusing on to achieve structures with thermodynamic stability, not multiple states for various folding conformations. In view of protein structure flexibility, many databases have cumulated information about protein or sequence regions involving intrinsically disordered protein (IDP).1616Lazar, T., Martınez-Pérez, E., Quaglia, F., Hatos, A., Chemes, L.B., Iserte, J.A., Méndez, N.A., Garrone, N.A., Saldaño, T.E., Marchetti, J., Velez Rueda, A.J., Bernadó, P., Blackledge, M., Cordeiro, T.N., Fagerberg, E., Forman-Kay, J.D., Fornasari, M.S., Gibson, T.J., Gomes, G-N.W., Gradinaru, C.C., Head-Gordon, T., Ringkjøbing Jensen, M., Lemke, E.A., Longhi, S., Marino-Buslje, C., Minervini, G., Mittag, T., Monzon, A.M., Pappu, R.V., Parisi, G., Ricard-Blum, S., Ruff, K.M., Salladini, E., Skepö, M., Svergun, D., Vallet, S.D., Varadi, M., Tompa, P., Tosatto, S.C.E., Piovesan D., PED in 2021: a major update of the Protein Ensemble Database for intrinsically disordered proteins,Nucleic Acids Research, Volume 49, Issue D1, (2021) D404–D411,1717Damiano Piovesan, Marco Necci, Nahuel Escobedo, Alexander Miguel Monzon, Andras Hatos ... Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D361–D367,1818Fukuchi, Satoshi et al. "IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature." Nucleic acids research vol. 40, Database issue (2012): D507-11.,1919Federica Quaglia, Balint Meszaros, Edoardo Salladini, Andras Hatos, Rita Pancsa ... DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation Nucleic Acids Research , Volume 50, Issue D1, 7 January 2022, Pages D480–D487,2020Damiano Piovesan, Marco Necci, Nahuel Escobedo, Alexander Miguel Monzon, Andras Hatos, Ivan Mičetić, Federica Quaglia, Lisanna Paladin, Pathmanaban Ramasamy, Zsuzsanna Dosztányi, Wim F Vranken, Norman E Davey, Gustavo Parisi, Monika Fuxreiter, Silvio C E Tosatto, MobiDB: intrinsically disordered proteins in 2021, *Nucleic Acids Research* , Volume 49, Issue D1, 8 January 2021, Pages D361–D367 The definitions of IDP are based on annotations of experimental data coming mainly from Nuclear Magnetic Resonance (NMR), Small-angle X-ray Scattering (SAXS) measurements and Molecular Dynamics (MD) simulations. However, an optimal approach for protein folding should obtain all possible folding conformations, expose folding difference between regions within a protein or between different proteins, including mutation or differentiation. Also, the most possible conformation and 3D structure should be extracted from a massive number of conformations.

Here, the protein structure fingerprint as novel approach to reveals the protein folding variations as well as the most possible conformation. A folden of element of 5 amino acid residues is firstly defined to probe the attribute of local folds, and then the local folds are extended to entire protein system to discover all possible folding conformations. First, a folden with 5 points connection as ball-and-stick is initiative model and make mathematical derivation. Without biological structure constrain, all folds are equivalently around each join point with topological uniformity, and all possible folds in geometric space form a complete and continuous aggregation. Second, the continuous aggregation of folding description is simplified by partition of space to reduce variable dimensions, and is applied to protein biological space. Then a set of 27 folding shapes are obtained which is able completely to cover various folding patterns for 5 successive amino acid residues. Third, with alphabetic description, these 27 folding shapes are represented by using 26 letters and "$" symbol. Thus, the topological model mathematically established the foundation to describe the protein backbone folding. As a set of 27 letters is applied to protein systems, it is called as the Protein Folding Shape Code (PFSC),2121Yang J, Comprehensive description of protein structures using protein folding shape code. Proteins;71.3:1497-1518 (2008). which essentially represent the folding shapes of 5 amino acid residues.

For protein with known 3D structure, its complete folding conformation can be described by PFSC string. The folding shape of any set of 5 successive amino acid residues is identified by a PFSC letter according the given coordinates of alpha carbon atoms. Along sequence from N-terminus to C-terminus, the conformation can be described by a string of PFSC letters. As one PFSC letter represents a folding shape of 5 successive amino acid residues, and two adjacent PFSC letters actually share partial of folding shape overlap of 4 amino acid residues. Thus any protein folding conformations can be completely described by a PFSC string, covering regular secondary structure fragments as well as irregular tertiary structure fragments.

For protein without given 3D structure, the comprehensive folding conformations for a protein are able to be exposed by local folding variations. In order to achieve this goal, all of possible permutations for 5 amino acids as well as all possible local folding shapes for each 5 amino acids are needed to well know. There are total 3,200,000 permutations for 5 amino acids based on 20 standard amino acids. For the permutations of 5 amino acids available in PDB, all folds have been first primarily collected. Then, for the permutations of 5 amino acids not available in PDB, their 3D structures were calculated by molecular dynamics simulations, and the folding shapes were obtained. Consequently, a new database 5AAPFSC, where the folding shapes of 5 amino acids are described by PFSC, is created to assemble all possible folding shapes for each permutations of 5 amino acid residues. A set of 5 amino acids may have one or more than one PFSC letters, but no more than 27 PFSC assignments. Each set of 5 amino acids may have different folding patterns and different number of PFSC letters. Therefore, according sequence only, all the possible folding shapes for each successive 5 amino acids from N-terminus to C-terminus can be thoroughly represented by continue sets of PSFC letters. The local folding variations are displayed in Protein Folding Variation Matrix (PFVM)2222Yang J. Protein Structure Fingerprint Technology. J Bioinform, Genomics, Proteomics X: 3(2): 1036, (2008). which the protein sequence is listed horizontally, and all folding shapes in the PFSC letters for each 5 successive amino acids are displayed vertically. The PFVM provides rich information to promote protein folding investigation. First, for a protein, the comprehensive local folding variations along sequence are simultaneously exhibited by PFVM. Second, the local folding variations are fluctuated with the folding pattern and number. Third, all possible conformations with an astronomical number for a protein can be assembled with various combinations of local folding variations, and the most possible conformation and 3D structure can be easily determined. Finally, the protein structure fingerprint produces the ensemble of conformations to probe the protein structures as well as the application in biological drug design and disease research.2323Yang J & Lee WH, Protein Structure Alphabetic Alignment, Protein Structure, Edited by Eshel Faraggi, InTech Publishers, (ISBN 978-953-51-0555-8), 133-156 (2012).,2424Yang J, Wu G, From Sequence to Protein Folding Variations. Biomedical Journal of Scientific & Technical Research, (2019).,2525Yang J, Zhang P, Cheng W X, et al. Exposing Structural Variations in SARS-CoV-2 Evolution. Scientifc Reports, 11:22042, (2021). Thus, the protein structure fingerprint provided a signification foundation for the solution of protein folding and applications.

METHODS

Protein Folding Shape Code (PFSC).

With protein folding fingerprint, the PFSC alphabetic string can provide a complete description for protein conformation. Mathematically, 5 points with successive connection in geometric space was firstly considered as a topological folding model. With derivation, the initial higher dimensions of topological folding space were reduced and the continuous space was partitioned, and then a set of 27 folding shapes was obtained which are able completely to cover various folding patterns for 5 points in sequential connection. These folding shapes can be representing with 27 letters including "$" symbol as a digitized expression. However, for biological protein, a set of 5 successive amino acid residues may not actually have all 27 folding shapes due to structural constrain. With alphabetical expression, the 27 folding shapes for 5 successive amino acid residues are called by Protein Folding Shape Code (PFSC) and displayed in the cubic of Figure 1. With integration feature, any PFSC letter in the cubic has partial folding similarity with its surrounding neighboring letters, and then a 27 PFSC may be transformed from one to each other by neighboring similarity. For example, the letter "A" represents a typical alpha-helix, and "H, D, V, Y, J and P" around "A" respectively has partial alpha-helical character in folds. In the same column with "A", "H" is an extensive helix fold and "D" is a compressive helix fold respectively. Nearby "A", "V, Y, J or P" has partial helical fold in N- or C-terminus respectively. The letter "B" represents a typical beta-strand, and "E, G, V, J, M and S" around "B" respectively has partial beta-strand in folds. Other remaining letters relate to irregular folds for tertiary folds, and also have partial similarity with neighboring letter in the cubic. Briefly, none of PFSC is isolated, and all 27 PFSC letters are formed into a meaningful ensemble with structural relevance. The blue arrows in Figure 1 indicated that the folding conformation for a protein with known 3D structure is able to be completely described by a PFSC string without gap. The PFSC letter is assigned to the folding shape of each 5 consecutive residues from N-terminus to C- terminus, and two PFSC letters next each other share folding shape of four amino acids. In summary, for any protein with given 3D structure, the PFSC alphabetical string is able to completely describe the folding conformation, covering secondary structure fragments as well as tertiary structure fragments.

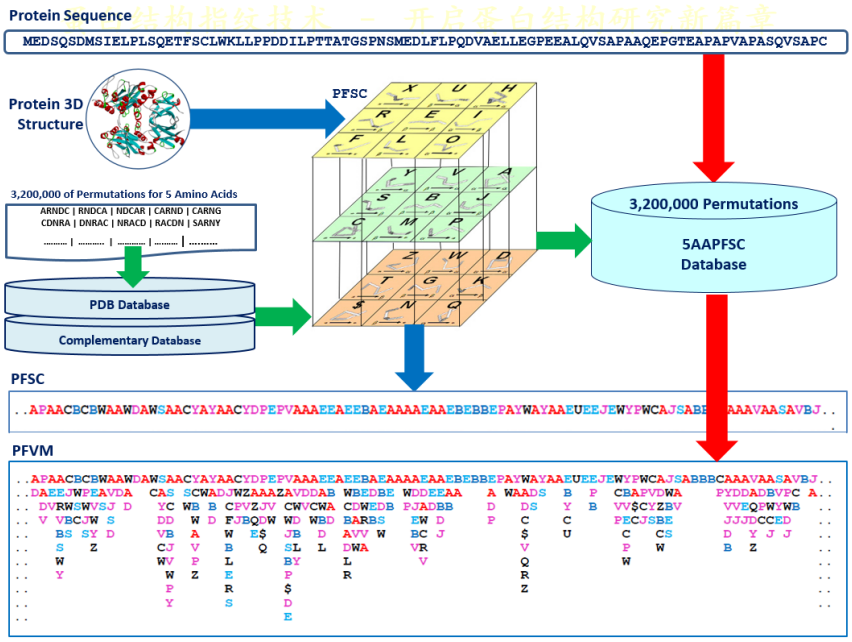


Figure 1. Protein folding shape code (PFSC) and protein folding variation matrix (PFVM). The green arrows indicated the process to construct 5AAPFSC database, which contained all folding shapes in PFSC letters for 3,200,000 of permutations of 5 amino acids. The blue arrows indicated the process how to obtain

the PFSC string from a protein with known 3D structure. The red arrows indicated the process how to obtain the PFVM from a protein sequence.

Protein Folding Variation Matrix (PFVM).

With protein folding fingerprint, the PFVM assembles the local folding variations along sequence, and it can construct an astronomical number of conformations while define the most possible conformation. Firstly, a database, which collects all possible folding shapes in PFSC letters, is created. Based on 20 of amino acids, there are 3,200,000 of permutations for 5 amino acids. All folding shapes for each permutations of 5 amino acids are collected from database or calculations. The folding shapes for most of permutations of 5 amino acids were firstly collected from 3D structural data in PDB. For the permutations of 5 amino acids do not exist in PDB, the folding shapes were computed by molecular dynamics simulation method and stored into a complementary database. Then, all folding shapes for 3,200,000 of permutations of 5 amino acids were converted into the PFSC alphabetic letters, and stored into a database named as 5AAPFSC. The procedure is indicated by green arrows in Figure 1. Actually, most of 5 amino acids have more than one folding shape, but the maximum number would not be more than 27. Each folding shape for 5 amino acids, however, have different weight according the frequency of appearance in PDB or the free energy for thermodynamic stability in results of computational simulation. Thus, the folding shapes for each set of 5 amino acids, which have higher frequency and lower free energy, are considered with most probability in folding conformation, and are assembled at the top rank in 5AAPFSC database.

According to protein sequence, the PFVM is constructed by extraction local folding variations from 5AAPFSC database. The local folding variations, which are represented by a set of PFSC letters for 5 successive amino acids, are extracted from 5AA-PFSC database and displayed in vertical column. Along sequence from N-terminus to C-terminus, the PFVM is formed. The procedure from a sequence to PFVM is indicated by red arrows in Figure 1. A diagrammatic sketch in Figure 2 illustrated in detail how the local folding variations were assigned according a protein sequence. The local folding shapes for each 5 consecutive amino acids along sequences are directly acquired from database 5AAPFSC. Starting from the first set of 5 amino acid residues at N-terminus, each step moves forward by one residue with sharing 4 amino acids from prior step until C-terminus. Therefore, the PFVM is generated to reveal the local folding variations for entire protein, the PFSC letters in each column show all possible local folding shapes for a set of 5 amino acids which are extracted from 5AAPFSC database. With PFVM, an astronomical number of PFSC strings can be constructed by taking any one PFSC letter from each column. Also, the most possible folding conformations can be predicted by the PFSC string which is consisted of the PFSC letter on top of each columns in PFVM. Furthermore, according PFSC string, the 3D structure of most possible conformation can be constructed as the predicted result. Therefore, the PFVM provides the significant information to study the protein fading problem.

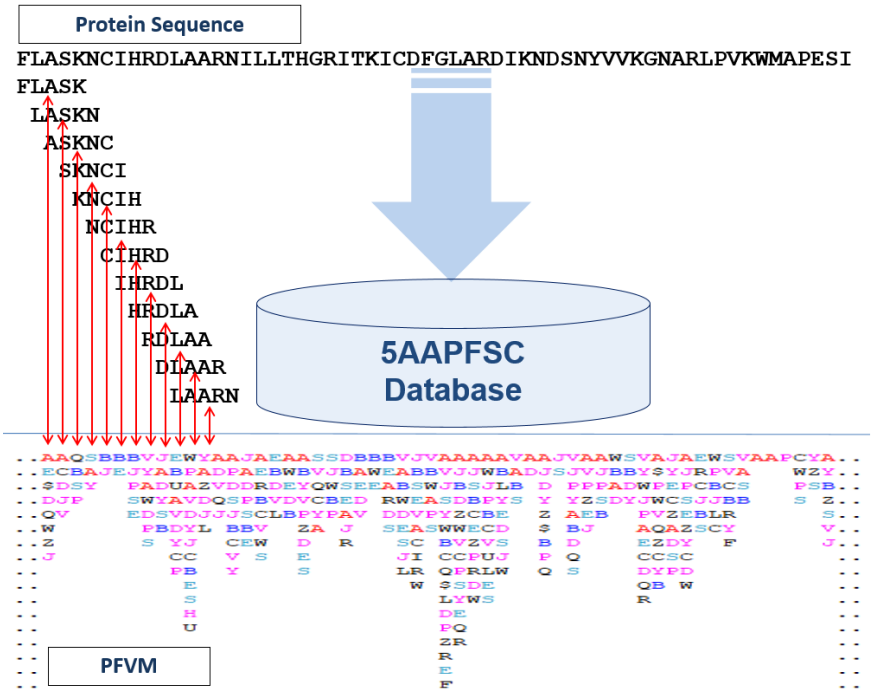


Figure 2. Acquirement of Protein Folding Variation Matrix (PFVM). The sequence is listed at the top horizontally. Then PFSC letters with color in column, which represent local folding shapes of 5 consecutive amino acids along sequences, are directly extracted from 5AAPFSC database. Starting from N-terminus, each step moves forward by one residue sharing 4 amino acids next each other. The PFSC letters for each 5 successive amino acids are vertically aligned down under the center of 5 amino acids. The PFSC letters are marked by colors: red is for typical helix fold; blue for typical beta fold; pink and light blue for folds with partial helix or beta; black for irregular folds.

All possible folding conformations.

Astronomical number of folding conformations for a protein can be constructed by the local folding variations in PFVM. To take one PFSC letter from each column in the PFVM is able to form one PFSC string, which is corresponding to one folding conformation. With fully using all of local folding variations from each column, all possible folding conformations will be constructed. Thus, the PFVM is one of the best optimized approach to stores all folding information for entire protein. The actual number of all possible folding conformations is able to be figured out. Under PFSC scheme, an astronomical number of all possible folding conformations for a protein is the product by multiplying the numbers of PFSC letters of each columns in PFVM.

Number of All Possible Conformations $= \prod_{i=1}^{n} m(i) = m(1) * m(2) * m(3) \ldots m(n)$

Where n is the number of columns in PFVM matrix and$m\ (i\ )$ is the number of PFSC letters in column i .

RESULTS

PFVM Assembling Local Folding Variations.

Based on protein sequence, the local folding variations of 5 successive amino acids can be directly obtained and exhibited in the PFVM. All PFSC letters for any 5 amino acids are extracted from the 5AAPFSC database, and vertically aligned down under the center of corresponding 5 amino acids. The order of PFSC letters in each vertical column is ranked according to the probability of occurrence from higher to lower and the free energy from lower to higher. Thus, the local folding variations for an entire protein can be

meaningfully displayed in PFVM. Two of human proteins, small ubiquitin-related modifier 1(SUMO1)and cellular tumor antigen p53 (P53), are taken as examples, and their PFVM will be acquired according the sequences respectively.

PFVM for SUMO1 protein

The protein of small ubiquitin-related modifier 1 (SUMO1)has highly flexible on the regions of N- and C-terminuses, which involves a variety of cellular processes, such as nuclear transport, transcriptional regulation, apoptosis and protein stability. The sequence of human protein SUMO1_HUMAN with 101 of amino acids is available from UniProt database. The PFVM was obtained within a few of seconds on a personal computer (4 x [Intel(R) Core(TM) i53337 CPU @1.80 GHZ], Windows 64bit operating system) after input of the sequence of SUMO1_HUMAN. The PFVM for SUMO1_HUMAN protein is shown in Table 1. It is obvious that each set of 5 amino acids has different folds. For example, starting from N-terminate, the first set of 5 amino acids "MSDQE" has 11 PFSC letters representing different folding shape. With moving forward by one amino acid, the second set of 5 amino acids "SDQEA" has 13 PFSC folding shapes; the third set of 5 amino acids "DQEAK" has 11 of different folding shapes; fourth set of 5 amino acids "QEAKP" has 9 different folding shapes; fifth set of 5 amino acids "EAKPS" has 8 different folding shapes and so on. Except the numbers different, the patterns of folding shapes for each set of 5 amino acids are different. With PFSC scheme, the distribution of the different numbers of folding shapes as well as the different types of folding shapes for entire protein of SUMO1_HUMAN can be simultaneously observed. Also, the pattern of folding variations of SUMO1_HUMAN in PFVM is overall agree with the disorder determined by PDB which is displayed on the bottom of Table 1, i.e. both N-terminus and C-terminus are more disorder than center region. Thus, the PFVM actually demonstrates the folding variations, and the local folding variations of protein SUMO1 relying on the order of amino acids in its sequence.

The PFVM in Table 1 also revealed the features of folding conformation for entire protein of SUMO1_-HUMAN. First, the PFVM is able to expose the flexibility or rigidity for protein folding conformations. In the middle portion of the sequence, the fragment (45-54), has a favor conformation as "AAAADAAAAA" with PFSC letters in first row which is almost typical alpha-helical conformation, and has the favor conformation as "DDDDADDDDD" in second row which is alike alpha-helical conformation. Also, the fragment in middle portion has fewer options changing folding variations than the fragments at both N-terminus and C-terminus. The revealed folding pattern in PFVM for SUMO1_HUMAN generally agrees with the knowledge from given protein 3D structural data in PDB. Seven of 3D structures for SUMO1_HUMAN protein are displayed in Table 2, which listed their 3D images with PDB ID, measurement methods, resolution, solvent and ligand, interacted protein, PH and temperature etc. Thus, these 3D structures may have different folding conformations. With comparison, the superimposition of 17 folding conformations from seven PDB 3D structures is displayed in Figure 3, where the fragments in middle region share common folding pattern while the fragments at both N-terminus and C-terminus are diverse. Second, the PFVM is able to provide the information to construct all possible conformations for a protein. Each conformation of SUMO1_HUMAN can be constructed by taking one PFSC letter from each column in PFVM, and is expressed by a PFSC string. Apparently, the conformations with astronomical number can be generated by various combinations of PFSC letters in PFVM, and its astronomical number can be calculated by multiplying the number of PFSC letters in each column for each 5 amino acids. For SUMO1_HUMAN, the astronomical number of all possible conformations is 1.587×10^{77}. Obviously, it is impossible to display all 3D folding structures of such huge number. Thus, the PFVM truly is an optimized mode to present the comprehensive local folding variations for entire protein, and it can generate all possible folding conformations for a protein. Furthermore, the most possible conformations with stable states can be constructed from PFVM. One of most probable conformations for SUMO1_HUMAN protein is directly consisted by PFSC letters on the first row in PFVM of Table 1. Therefore, the PFVM provides significant folding information for SUMO1_HUMAN.

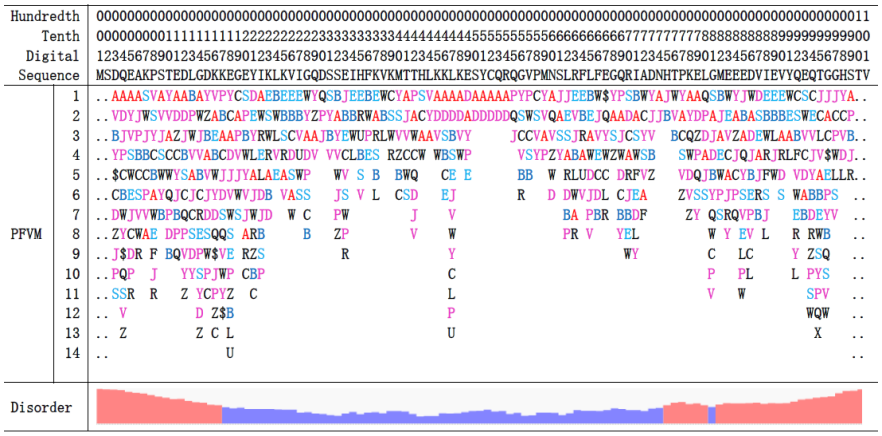


Table 1. The Protein Folding Variation Matrix (PFVM) for Human small ubiquitin-related modifier 1 (SUMO1_HUMAN) protein. Top section: the sequence of 101 of amino acids of SUMO1_HUMAN with numeric ruler. Section below sequence: the PFVM. The PFSC letters in each vertical column represent the local folding variation for 5 successive amino acids along sequence. The order of PFSC letters in each vertical column is ranked from higher to lower according the frequency numbers of folding shapes in PDB. The PFSC letters are marked by colors: red is for typical helix fold; blue for typical beta fold; pink and light blue for folds with partial helix or beta; black for irregular folds. Bottom section showed the disorder determined by PDB.

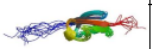
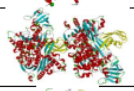
PDB ID	Method	Resolution (Å)	Model/chain	Positions	Image	Solvent and Ligand	Interacted Protein	pH	Temperature (K)
1A5R	NMR		1-10	1-101		H2O		7.0	300
3KYD	X-ray	2.61	D	1-94		5'-{[(3-aminopropyl)sulfonyl]amino}-5'-deoxyadenosine; ETHYLENE GLYCOL; H2O	SAE1_HUMAN;		279
1WYW	X-ray	2.10	B	1-97		MAGNESIUM ION;SODIUM ION; H2O	TDG_HUMAN	8.5	293
1Y8R	X-ray	2.75	C/F	1-97		H2O	SAE1_HUMAN;	5.6	278
2PE6	X-ray	2.40	B	1-97		H2O	UBC9_HUMAN	8.5	291
3KYC	X-ray	2.45	D	1-97		5'-deoxy-5'-(sulfamoylamino)adenosine	SAE1_HUMAN;	6.4	291
2BF8	X-ray	2.30	B	21-97		H2O	UBE2K_BOVIN	8.5	

Table 2. Experimental and environmental parameters for seven protein structures of (SUMO1_HUMAN) protein. Seven 3D structures are obtained from PDB which were determined by NMR measurement or crystal X-ray diffraction. Each structure was measured under different conditions, such as method, resolution, solvent and ligand, protein interacted, PH and temperature etc. The SUMO1 protein is marked in the color yellow in images of structure complexes. The model index, chain name and fragment position in sequence are indicated.

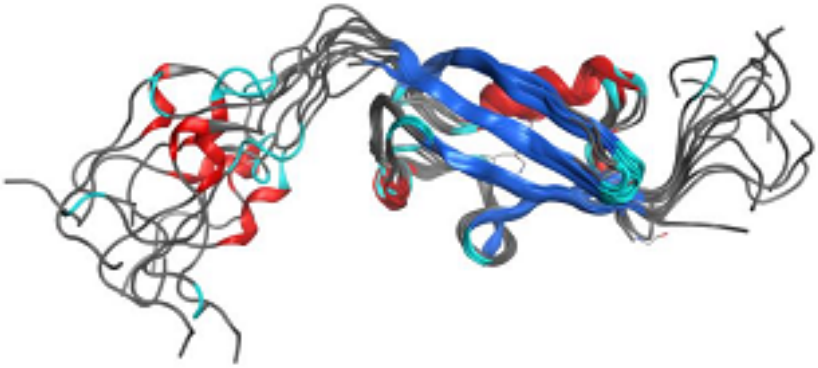
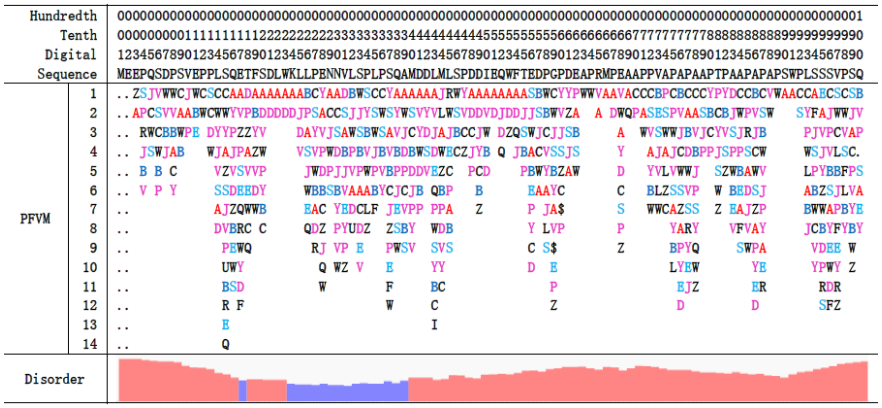


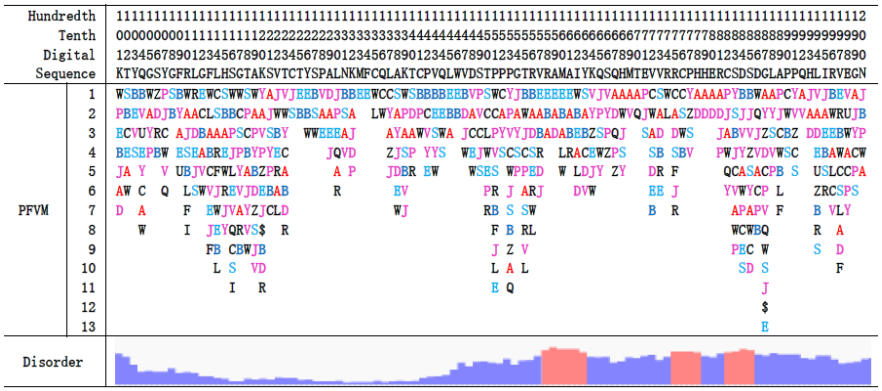
Figure 3. The superimposition of conformations for SUMO1_HUMAN. Seven structures in PDB provide 17 conformations. 1A5R has ten conformations, 1Y8R has two conformations and 3KYD, 1WYW, 2PE6, 3KYC and 2BF8 has one conformation individually. The ribbons are colored by segments as the followings, black for loop, red for helix, blue for strand and light blue for turn.

PFVM for P53 protein

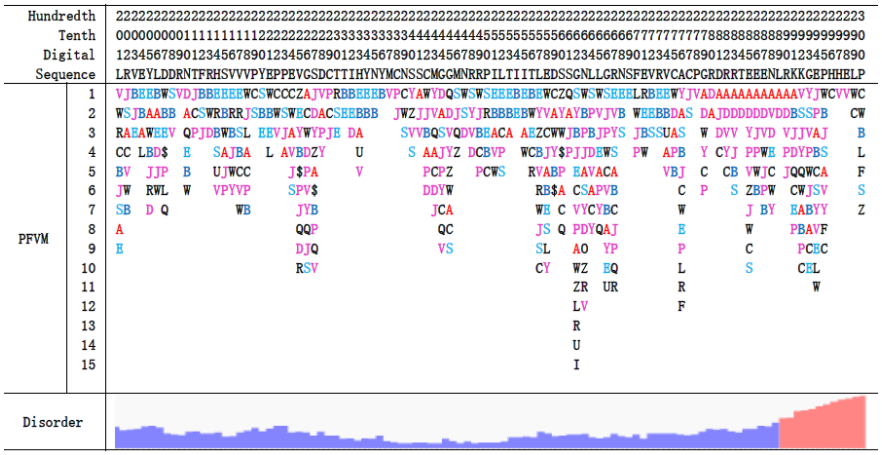
The protein of Human Cellular tumor antigen p53(P53_HUMAN)is another sample to illustrate how the local folding variations in PFVM correlate with the order of amino acid along sequence. The p53 involves many mechanisms for anticancer functions and plays a role in apoptosis, genomic stability, and inhibition of angiogenesis. The human TP53 gene encodes 9 of protein isoforms11Surget S, Khoury MP, Bourdon JC, "Uncovering the role of p53 splice variants in human malignancy: a clinical perspective". OncoTargets and Therapy. 7: 57–68, (2013)., and the canonical sequence of isoform p53-alpha with 393 amino acids is available in UniProt database. According the sequence, the PFVM of P53_HUMAN protein is obtained and shown in Table 3. Although the protein of p53-alpha for human has 393 amino acids in sequence, the PFVM of p53-alpha was acquired within one minute using a personal computer. First, the PFVM directly displayed the fluctuation of local folding variations along the sequence. For examples, some sets of 5 successive amino acids have only one folding shape representing with one PFSC letter in PFVM, such as fragments in sequence of "DEAPR" (61-65), "EAPRM" (62-66), "APSWP" (88-92), "PSWPL" (89-93), "NKMFC" (131-135) and "KMFCQ" (132-136) etc. Some sets of 5 successive amino acids have more folding shapes representing with higher number of PFSC letters in PFVM, such as "PLSQE" (13-17) has 14 folding shapes; "SDGLA" (185-189) has 13; "SSGNL" (260-264) 15 and "SKKGQ" (371-375) 16 etc. Second, the pattern of folding variations of P53_HUMAN in PFVM showed many regions agree with the disorder determined by PDB which is displayed on the bottom of Table 3. For examples, regions of 20-25, 124-136, 230-241 and 332-347 have less variations and disorder. However, the PFVM may exposed the folding variations more detail. Third, all possible conformations for the p53-alpha protein can be constructed by various combinations with taking one PFSC letter from each column in PFVM. Actually, with PFVM, an astronomical number of folding conformations for the p53-alpha protein can be constructed and represented by $2.008\text{x}10^{270}$ of PFSC strings. It showed that the astronomical number of all possible conformations for p53-alpha protein were not vague anymore because the number can be explicitly calculated based on PFVM.



(Continue)



(Continue)



(Continue)

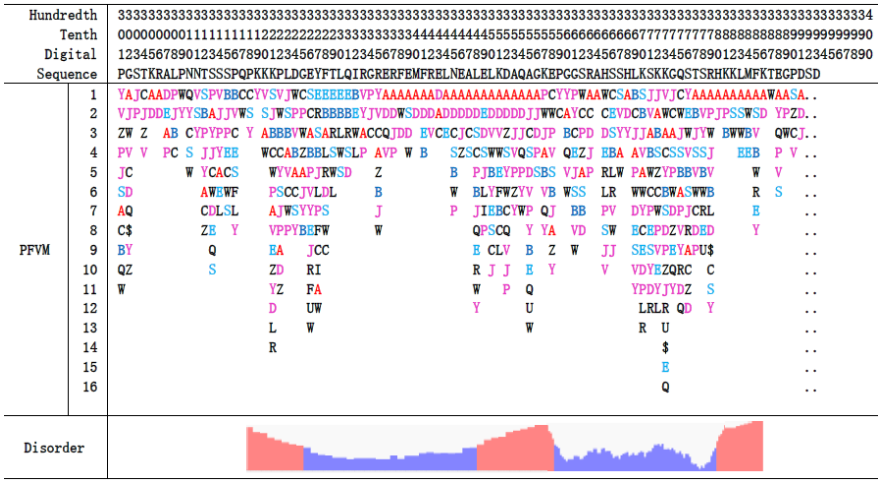


Table 3. Protein Folding Variation Matrix (PFVM) for protein of Human cellular tumor antigen p53 (P53_HUMAN). Top section: the sequence of 393 of amino acids of P53_HUMAN with numeric ruler. Section below sequence: the PFVM. The PFSC letters in each vertical column represent the local folding variation for 5 successive amino acids along sequence. The order of PFSC letters in each vertical column is ranked from higher to lower according the frequency numbers of folding shapes in PDB. The PFSC letters are marked by colors: red is for typical helix fold; blue for typical beta fold; pink and light blue for folds with partial helix or beta; black for irregular folds. Bottom section showed the disorder determined by PDB.

The PFVM has capability to discover the change of local folding conformation caused by mutation, even if 3D structural data is lacked. The P53 has an important role as a tumor suppressor, but the mutations severely compromise the tumor suppression. The mutations may happen in various positions or make different substitution in same position of sequence related to different diseases and biological functions. The different substitution of amino acid residue 183 of P53, for example, related to different biological functions. The mutation S183E abolishes phosphorylation, but the mutation S183A inhibits its transcriptional activity.[17] The replacement of amino acid residue triggered the changes of folding conformation. The changes of local folding shapes around residue 183 of P53_HUMAN are revealed by PFVM and the results are exhibited in Table 4. The mutation on a residue primarily affects around five columns of local folding variations in PFVM which are highlighted with a yellow color. It is apparent that with comparison of native state, the folding patterns are changed and the numbers of folding variations are reduced respectively after mutating S183E and S183A in P53 sequence. The change of conformation, which is caused by a residue replacement in mutation, is hard to be observed by experimental measurements, such as NMR, X-ray crystallography or Cryo-electron microscopy, and it is hard to be compared by the results from computational molecule dynamics simulation. However, the PFVM provided a useful means to reveal the difference of folding conformation between before and after mutation. Additionally, the folding conformation changes in PFVM caused by mutation are further demonstrated that the protein folding was associated with the order of amino acids in sequence.

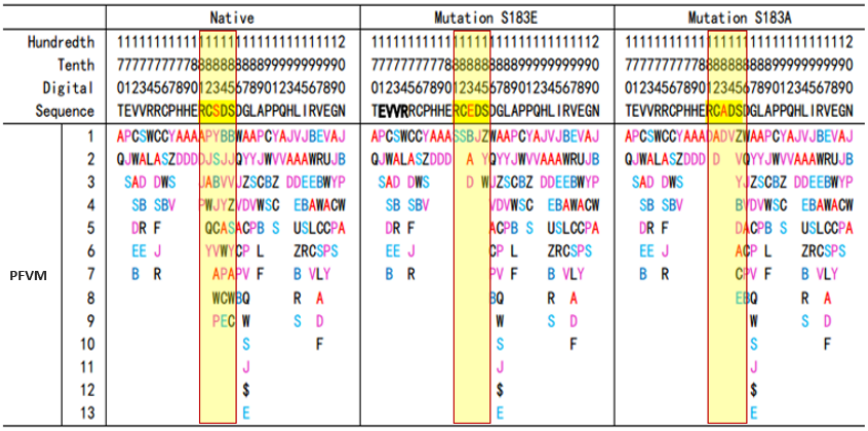


Table 4. Comparison of Protein Folding Variation Matrix (PFVM) for mutation on residue 183 of P53_HUMAN. Top section: the sequence fragment (170-200) of P53_HUMAN with numeric ruler. Bottom section: the local folding variation in PFSC letters. The PFSC letters in each vertical column represent the local folding variation for 5 successive amino acids in sequence. The order of PFSC letters in each vertical column is ranked from higher to lower according the frequency numbers of folding shapes in PDB. The PFSC letters are marked by colors: red is for typical helix fold; blue for typical beta fold; pink and light blue for folds with partial helix or beta; black for irregular folds.

PFVM Embracing Known Structures

The PFVM can generate all possible folding confirmations for a protein with an astronomical number, of course, it should embrace the protein conformations of known 3D structures. For a protein, it usually may have multiple structural data in PDB, but these 3D structures in some degrees have difference in folding conformations because of different measurement methods and environments. With PFVM, different conformations in PFSC strings can be constructed by the PFSC letters from each column in matrix, and the conformations for protein with known 3D structures should be one of these PFSC strings.

SUMO1 known 3D structure conformations covered by PFVM

The SUMO1_HUMAN protein has more than 50 of 3D structures available in PDB, and it is not surprised that these structures have both similarity and difference in folding conformations. Here, seven of 3D structures of SUMO1_HUMAN (Table 2) as examples are taken to compare the folding conformations. First, these 3D structures were obtained by different measurement methods, under different environments and interaction with different molecules. Second, many conformations can be generated from these protein 3D structures. 1A5R has 10 of conformations which is NMR experimental data in water solution.11Bayer, P., Arndt, A., Metzger, S., Mahajan, R., Melchior, F., Jaenicke, R., Becker, J. Structure determination of the small ubiquitin-related modifier SUMO-1 [J]. Journal of Molecular Biology, 280(2):275, (1998). The rest of protein complex structures were measured by X-ray crystal diffraction, 1Y8R has two chains and 3KYD, 1WYW, , 2PE6, 3KYC and 2BF8 has one chain respectively22Olsen SK1, Capili AD, Lu X, Tan DS, Lima CD.,Active site remodelling accompanies thioester bond formation in the SUMO E1, Nature. 463(7283):906-12, (2010).·33Baba D, Maita N, Jee J G, et al. Crystal structure of thymine DNA glycosylase conjugated to SUMO-1[J]. Nature, 435(7044):979-82, (2005).·44Lois L M, Lima C D. Structures of the SUMO E1 provide mechanistic insights into SUMO activation and E2 recruitment to E1[J]. Embo Journal, 24(3):439-51, (2005).·55Capili A D, Lima C D. Structure and Analysis of a Complex between SUMO and Ubc9 Illustrates Features of a Conserved E2-Ubl Interaction[J]. Journal of Molecular Biology, 369(3):608-618, (2007).·66Pichler A, Knipscheer P, Oberhofer E, et al. SUMO modification of the ubiquitin-conjugating enzyme E2-25K[J]. Nature Structural & Molecular Biology, 12(3):264-9, (2005).. Thus, these seven SUMO1 protein structures contain total of 17 folding conformations, and they are respectively converted into 17 of

PFSC strings according their coordinates of alpha C-atoms. The PFSC strings for fragment (21-84) for 17 folding conformations of SUMO1_HUMAN are aligned and listed in the top section of Table 5.

It is easy to compare the similarity or dissimilarity of structure conformations with alignment of PFSC alphabetical strings. We try to find out how many of different types of local folding shapes in each column exist for 17 of folding conformations for seven 3D structures. In order to do so, all PFSC letters for conformation of 1A5R-01 (model 1) are firstly counted and marked by yellow background color. Then, for the rest of 16 PFSC strings, any PFSC letter in same column not same as 1A5R-01 or other will be highlighted by yellow. Thus, the PFSC letters with yellow color on each column will reveal what different types of folding shapes exist for each of 5 successive amino acid residues. For the 17 known conformations, the fragment (21-84) in SUMO1 has 64 columns, i.e. 64 sets of 5 successive amino acid residues in structure. There are 9 columns which have the identical local folding shapes as 1A5R-01, and the remainder 55 columns have at least one different local folding shape from 1A5R-01. Therefore, it is not too hard to detect the similarity or dissimilarity for 17 conformations with alignment of PFSC strings. First, all 17 conformations have similar secondary structures distributed along sequence which is observed by the PFSC letter colors. Second, none of 16 conformations are matching 1A5R-01. Third, it is apparent that each structure has unique folding conformations and can be distinguished from each other. Specifically all six structures from X-ray crystallography have larger alterations of local folds than 1A5R-01 conformations from NMR measurement. In summary, despite overall similarity of secondary structure, the known 3D structures of SUMO1 protein have different folding conformations which are well revealed by alignment of PFSC strings.

The PFVM of SUMO1_HUMAN protein contains the comprehensive folding information which embraces all 17 folding conformations of known structures. The PFVM for SUMO1_HUMAN protein is displayed on the bottom section of Table 5. The PFSC letters on each column represent the possible folds for each 5 successive amino acids. In order to show the PFVM covering the folds in the known structures, any PFSC letters in PFVM, which already have appeared on same column for known structures, are marked by yellow. It is apparently that all folding letters with yellow in the known structures are enclosed by the PFVM. For example, the PFVM in column 38 has 3 types of folds (W, B and L) covering the folds of "W" and "L" in given structures; the PFVM in column 39 has 4 types of folds (C, S, W and R) covering the folds of "C", "S" and "R" in given structures; the PFVM in column 40 has 6 types of folds (Y, S, V, Z, B and C) covering the folds of "V", "B" and "Y" in the given structures and etc. Thus, the PFVM has complete local folding variations, which is able to cover the folding changes in given 3D structures of SUMO1 protein. Furthermore, the most possible conformations are able to be predicted by PFVM. The PFSC string on first row of PFVM in Table 5, which is comprised of the local folding shapes with the most tendencies, directly present one of the most possible conformations for SUMO1 protein. With yellow colors of PFSC letters, it is apparent that the predicted conformation is overall aligned well with all of 17 conformations from known 3D structural. Also, the most possible conformation has 60 among 64 PFSC letters with the marked yellow color, which indicated the most possible conformations is matched with the known structures. In conclusion, the PFSC string on first row in PFVM provides one of well predicted conformations. Also, comprehensive local folding variations in PFVM are able to cover the various conformations of SUMO1 protein from known 3D structures.

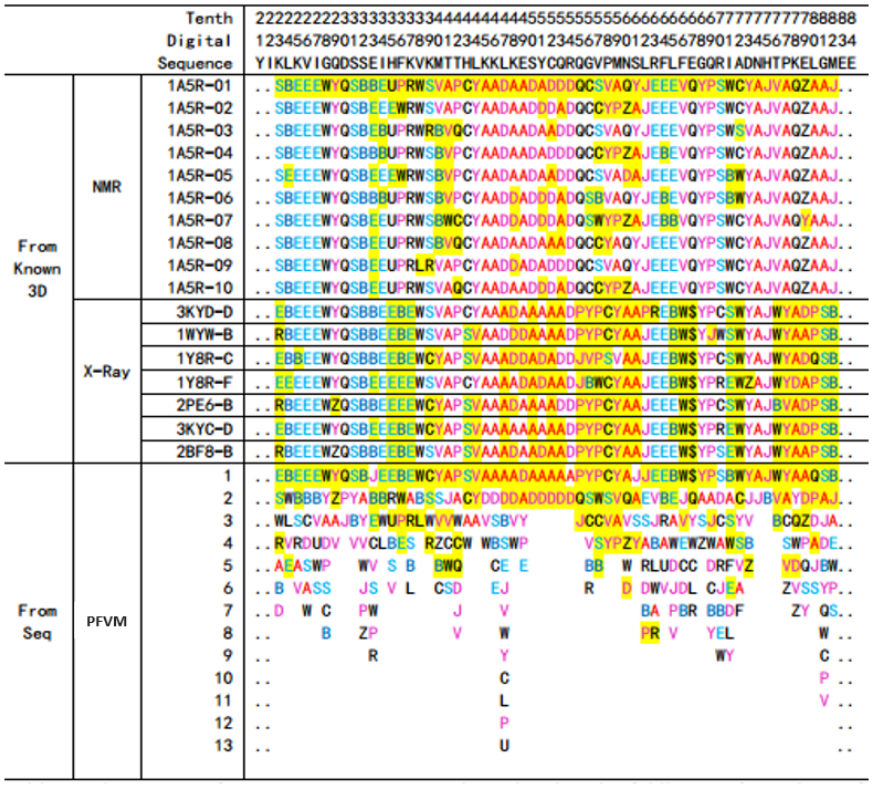


Table 5. The PFVM of SUMO1_HUMAN embracing folding conformations of protein with known 3D structures. Top section: the sequence (21-84) of SUMO1_HUMAN protein with numeric ruler. Middle section: the 17 PSCF strings for conformations of 7 known 3D structures of SUMO1_HUMAN. Seven of structures are available in PDB: 1A5R is NMR data with 10 models; 3KYD, 1WYW, 1Y8R, 2PE6, 3KYC and 2BF8 are crystal X-ray diffraction data with one chain respectively and 1Y8R with two chains. The all PFSC letters of 1A5R-01 is first marked by yellow color. Other PFSC letters in same column for these given structures are marked by yellow color if they differ from 1A5R-01. Bottom section: the local folding variation with PFSC letters for PFVM of SUMO1_HUMAN protein. The PFVM of SUMO1_HUMAN protein. The PFSC letters in each column in PFVM are highlighted by yellow color if the corresponding local folding shapes for 5 successive amino acids in given 3D structures are yellow. The PFSC letters are marked by colors: red is for typical helix fold; blue for typical beta fold; pink and light blue for folds with partial helix or beta; black for irregular folds.

P53 known 3D structure conformations covered by PFVM

Many P53_HUMAN structures are available in PDB from X-ray crystallization measurements. Twelve of protein 3D structures of P53 (4AGP, 2YBG, 2XWR, 2X0U, 2J1X, 2J1Y, 2FEJ, 2BIN, 3D05, 3D06, 3ZME and 5LAP) are displayed in Figure 4. The PSCF strings as folding conformations are obtained directly according 3D structures of regions (101-200) in sequence of 12 given P53 structures, and are aligned and displayed on the top section of Table 6. Also, the PFVM as local folding variations is obtained according merely the sequence (101-200) of P53 and 5AAPFSC database, and is displayed on the bottom section.

The PFVM is able to cover all conformations of given 3D structures of P53 protein. In order to verify what types of local folding shapes occur in these given P53 structures, all PFSC letters in 2YBG are first highlighted by yellow as starting point in Table 6, and then the PFSC letters for other 3D structures are checked down each column against 2YBG. If the PFSC letters are not covered by 2YBG, the letters will be highlighted by yellow. Thus, all local folding variations existing in these given structures are shown by

yellow. Furthermore, to find out whether the folding conformations of 12 given structures are enclosed by PFVM, the PFSC letters in each column in PFVM are marked by yellow if the PFSC letter in same column appears in given 12 structures. Therefore, it is easy to observe the local folding conformations of 12 given structures are indeed enclosed by PFVM. For examples, the PFVM in column 104 provides 8 types of folding shapes (B, V, U, E, Y, C, A and W) covering the folds of "B" and "V" in the given structures; the PFVM in column 105 provides 4 types of folding shapes (W, A, Y and P) covering the folds of "W" and "A" in the given structures; the PFVM in column 120 provides 11 types of folding shapes (Y, A, V, P, Z, E, J, $, B, D and R) covering the folds of "Y" and "Z" in the given structures; the PFVM in column 139 provides 7 types of folding shapes (W, P, A, S, B, V and J) covering the folds of "P" and "W" in the given structures; the PFVM in column 199 provides 6 types of folding shapes (A, J, Y, C, P and S) covering the folds of "A", "Y" and "J" in the given structures and etc. In conclusion, the PFVM is able to cover the local folding shapes of 12 given 3D structures for P53 protein, and the PFVM reserves far more local folding variations for P53 protein in other environments.

The PFVM provides the information to predict the possible fording conformation for protein. Despite a massive number of folding conformations for P53 protein can be generated with the combination of various PFSC letters in PFVM, the PFSC string at the first row in PFVM, which is named as PFVM-01, represents one of the most possible conformations for P53 protein. The PFVM-01 can be compared with the conformations of 12 given 3D structures in Table 6. Overall, with observation in color, the secondary structure fragments of PFVM-01 conformation are well aligned with all 12 conformations of given 3D structures of P53 protein. Also, 99 among 100 PFSC letters in the PFVM-01 conformation appear in the 12 given 3D conformations of P53 protein. Therefore, the PFSC string in the first row of PFVM provides one of best conformations for P53 protein prediction.

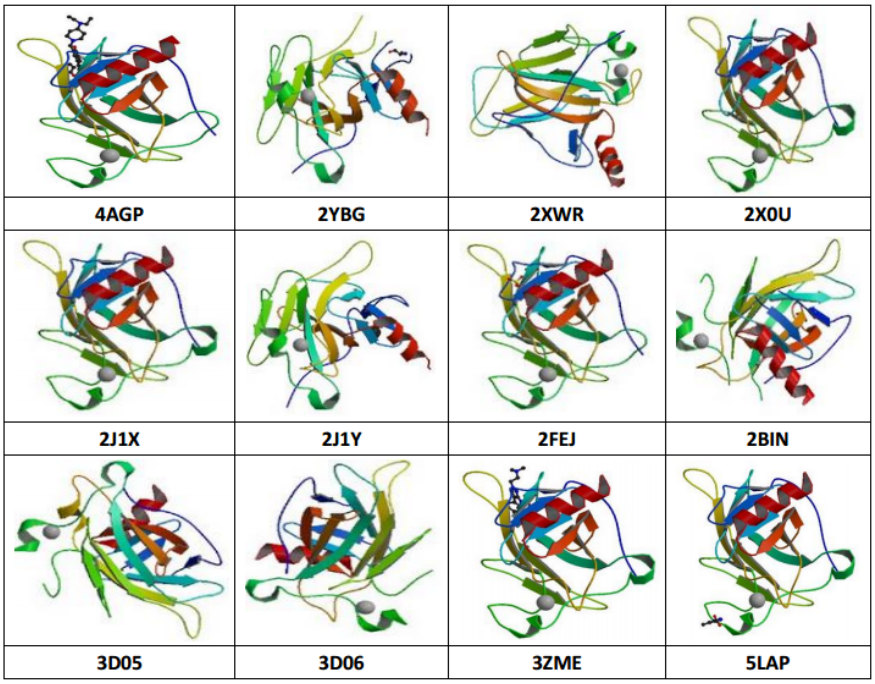


Figure 4. 3D structure images of folding conformations of 12 structures for P53_HUMAN protein displayed with solid ribbon style. The structure data was obtained from PDB and the PDB ID is listed below the image.

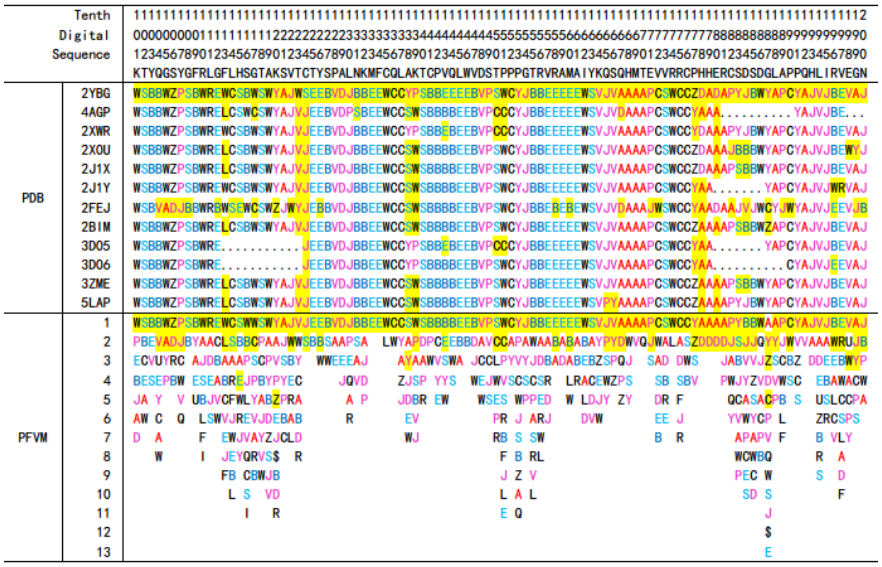


Table 6. The PFVM of P53_HUMAN embracing folding conformations of known 3D structures. Top section: the sequence (101-200) of P53_HUMAN protein with numeric ruler. Middle section: the PSCF strings for 12 known 3D structures of P53_HUMAN which are available in PDB: 2YBG, 4AGP, 2XWR, 2X0U, 2J1X, 2J1Y, 2FEJ, 2BIM, 3D05, 3D06, 3ZME and 5LAP. All PFSC letters for 2YBG is first marked by yellow color. Other PFSC letters in same column for these given 3D structures are marked by yellow color if they differ from 2YBG. Some pieces of missed fragments without atomic coordinate data in structure database in PDB are indicated by dots. Bottom section: the PFVM of P53_HUMAN protein. The PFSC letters in each column in PFVM are highlighted by yellow color if the corresponding local folding shapes for 5 successive amino acids in given 3D structures are yellow. The PFSC letters are marked by colors: red is for typical helix fold; blue for typical beta fold; pink and light blue for folds with partial helix or beta; black for irregular folds.

Astronomical Number of Conformations

The number of total conformations for a protein can be figured out according the aggregation of local folding variations in the PFVM. Generally known, total number of possible folding conformations for protein is large, but the so called astronomical number of conformations for any protein is blurred. However, with PFVM, the number of total conformations for a protein can be explicitly calculated based on the PFVM because total number of conformations for a protein is the product of numbers of PFSC letters at each column in PFVM. The total numbers of all possible conformations for ten proteins as samples are listed in Table 7. These samples represent a wider variety of proteins and the lengths of sequences with a range from 101 to 1,382 of amino acids. The results show that the total numbers of possible conformations for these proteins are large which have a scope from 10^{77} to 10^{784}. Two proteins among these samples, PDCD1_MOUSE and PDCD1_HUMAN, belong to same gene code PDCD1 with equal length of sequence, but they are respectively for human and mouse species. Despite both proteins having the equal length of sequence of 288 amino acids, the alignment revealed 172 of amino acids identical (about 60%) between two protein sequences. According PFVM of both proteins, total numbers of conformations respectively are 3.028×10^{151} and 2.827×10^{147} because of the protein differentiation in sequence for different organisms. The results demonstrated that the total number of conformations of protein related to the length of sequence as well as the composition of amino acids in sequence. In summary, based on protein sequence, the PFVM does not only expose the comprehensive local folding variations and predicted the most possible conformations, but also provided the actual number of total conformations for a protein.

UniProt Entry	Protein names	Biology Function & Disease Related	AA Length	Total Number
SUMO1_HUMAN	Small ubiquitin-related modifier 1	Cellular processes, such as nuclear transport, transcriptional regulation, apoptosis, and protein stability	101	1.587×10^{77}
INS_HUMAN	Insulin	Insulin decreases blood glucose concentration;Diabetes mellitus	110	4.940×10^{96}
SYUA_HUMAN	Alpha-synuclein	Regulation of dopamine release and transport; Induces fibrillization of microtubule-associated protein tau; Parkinson disease	140	2.385×10^{126}
TNFA_HUMAN	Tumor necrosis factor	Induce cell death of certain tumor cell lines; Impairs regulatory T-cells function; Immune system diseases; Rheumatoid arthritis	233	3.028×10^{151}
PDCD1_MOUSE	Programmed cell death protein 1	Inhibitory cell surface receptor involved in the regulation of T-cell function during immunity and tolerance	288	2.827×10^{147}
PDCD1_HUMAN	Programmed cell death protein 1	Inhibitory cell surface receptor involved in the regulation of T-cell function during immunity and tolerance	288	1.762×10^{155}
P53_HUMAN	Cellular tumor antigen p53	A tumor suppressor in many tumor types; induces growth arrest or apoptos; Cancer.	393	2.008×10^{270}

UniProt Entry	Protein names	Biology Function & Disease Related	AA Length	Total Number
A4_HUMAN	Amyloid-beta A4 protein	neurons relevant to neurite growth, neuronal adhesion and axonogenesis; Alzheimer disease	770	$1.875x10^{460}$
EGFR_HUMAN	Epidermal growth factor receptor	Receptor tyrosine kinase binding ligands of the EGF family and activating several signaling cascades to convert extracellular cues into appropriate cellular responses; Non-small cell lung cancer	1,210	$5.693x10^{743}$
INSR_HUMAN	Insulin receptor	Receptor tyrosine kinase which mediates the pleiotropic actions of insulin; Diabetes mellitus	1,382	$2.232x10^{784}$

Table 7. Total number of possible folding conformations for ten proteins respectively. The protein names with other information were obtained from UniProt database. Total number: the number of folding conformations was obtained according PFVM. The PFVM for these ten proteins are listed in the supplemental file.

Prediction of Most Possible Conformation and 3D Structure

The most possible conformation and 3D structure for protein can be predicted from its PFVM. With a protein sequence, the local folding variations are collected in PFVM. For examples, the PFVM of SUMO1_HUMAN is at Table 1; the PFVM of P53_HUMAN at Table 3 and the PFVM of K4GSD6_9SAUR, C4IXC1_-9TELE, A0A851ZE52_9AVES and EP3B_HUMAN in supplementary document. The alphabetic PFSC string on top of PFVM, which is named as PFVM-01, represents the most possible folding conformation. Their PFVM-01 are listed on Table 8, which are the most possible folding conformations for proteins. In PFVM-01, each of PFSC letter represents the folding shape of 5 amino acids in sequence, two PFSC letters next each other share 4 amino acids, and then each PFVM-01 is a PFSC string for folding conformation from N-terminus to C-terminus. As the PFSC letters in PFVM-01 are on top folding shapes in PFVM, the PFVM-01 represents the most possible conformation for a protein.

UniProtKB Protein Name	PFVM-01
SUMO1_HUMAN Small ubiquitin-related modifier 1	AAAASVAYAABAYVPYCSDAEBEEEWYQSBJEEBEWCYAPSVAAAADAAAAAPYPCYAJJEEBW$YPSBWYA JWYAAQSBWYJWDEEEWCSCJJJYA
P53_HUMAN Cellular tumor antigen p53	ZSJVWVAAWCCCCCYADAAAAAAAABCYADDJWSCVYAAAAAVAJSYAAAAAAAASSBCYYPWWVJAVAWCSEP CBCCCSPYDCCBCJLAPCYAECSCSBWSBBWZPSBWRELCSBWSWYAJVJEEBVDJBBEEWCCSWSBBBBEE BVPSWCYJBBEEEEEWSVJVAAAAPCSWCCYAAAAPSBBWAAPCYAJVJBEVAJVJBEEBWSVDJBWREEEW CSWCCCZAJVPREBEEEBVPCYJWYDQSWSWSEEEBEBEWCZQSWSWSEEELRBEEWYJVADAAAAAAAAAA ACSJBCVWWCYAWYAADPYBVAPVJBCCAVSVJWCSEEEEEBVPYAAAAAAADAAAAAAAAAAAAAPCYYPW AAWCSAABJAVJJYAAAAAAAAAAWCASB
K4GSD6_9SAUR 35 G protein-coupled receptor 149	BSAEDAEESAAADAAADYABAPVPAAAAEEBCBBYABAAWAEAESAEAWBAAWSASQSBAAWCCEJADABEA BAADAADABBWAVJSJELJBCCAEPEAWPCAPAAAAAAADEBDAAAWDBAAABECBEJAELAYWAVVVEBBA AAAEVJPZDJQSSWSVVAAAVBJCSVVADWAEVAAVVVVYJAAEEEEZYYYASAADDVVSPCVWAVRBEEEA AAWAQEEAABA
C4IXC1_9TELE Cytochrome c oxidase subunit 2	PYAAQCWAAAYJEDAAADBRYAAAABBDAASDAAWEAAABEJEAJVEDBCAWACCPAAAEEADAAAAAABA AAAAACAASAAAAAWPSBJVAPAEWJWPZCZPREEBBVJZAAAAACEERWWYARBWYAJEAAAJBEAAAVW AWDBSAEBEBPYAAJSBEEBADAESBBBAWWYAJEAASWAADJWRSCEDEWSVPSWFZBSAAJBRAAAAWA SAADUESAAAPAA
A0A851ZE52_9AVES ALP1 protein	DASAABAAASVAABVVAAABCSYABJVAYAAAAWSAABABAAEAJAJBADRDADADWWAEAAAQADAWAAAW AAEBEAVAAAAABEDSDAJSSAAAAAEPJPCAAVCBPPPCCCADAAQBPZAAAAAEWWWWAWEZA
EP3B_HUMAN Epididymal secretory protein E3-beta	VAJEDAAAAAAABAABDAEAAABAEEAAAAAADDEDAWVAVLAYAEESESAADSAAAAAAVWJJEWEEEEEEE ESVDAEEBSAAAADJWSEBWBBJPCBAAABEAAABAVABAEEBYALBPEJJEAPYDQCWAJAAAAWAAWEV

Table 8. The most possible folding conformations of PFVM-01. Left column listed the UniprotKB and protein name; right column listed the PFVM-01 which is the PFSC string on top of PFVM. The PFVM of SUMO1_HUMAN is at Table 1; the PFVM of P53_HUMAN at Table 3 and the PFVM of K4GSD6_-9SAUR, C4IXC1_9TELE, A0A851ZE52_9AVES and EP3B_HUMAN in supplementary document.

With PFVM-01, the protein 3D structure can be constructed which represents the predicted structure. The PFVM-01 is a PFSC string, which represented the folding conformation for a protein. First, with searching for similar conformations, an initiative 3D structure can be constructed according an entire PFVM-01. Using high throughput screening PDB, the similar structural conformations can be obtained according folding similarity score with comparison between the PFVM-01 and all PFSC strings in database. Thus, the initiative 3D structure can be constructed according the 3D structure with highest folding similarity score. Second, the parts of PFVM-01, particularly tertiary fragments, can be further searched to improve 3D structure. Third, the side chains are added for each residue. Finally, the constructed 3D structure is optimized by computed for free energy minimization. With PFVM-01, six of 3D structures are predicted and displayed in Figure 5. First row displays the predicted 3D Structure from PFVM-01 of SUMO1_HUMAN; second row displays the predicted 3D Structure from PFVM-01of P53_HUMAN. The SUMO1_HUMAN and P53_HUMAN have the known 3D structures, so the comparison between known 3D structure and the predicted structure (in brown color) are shown on left side. Although proteins of K4GSD6_9SAUR, C4IXC1_9TELE, A0A851ZE52_9AVES and EP3B_HUMAN do not have 3D structures available in PDB, their most possible 3D structures can be predicted by PFVM-01 are displayed respectively in Figure 5.

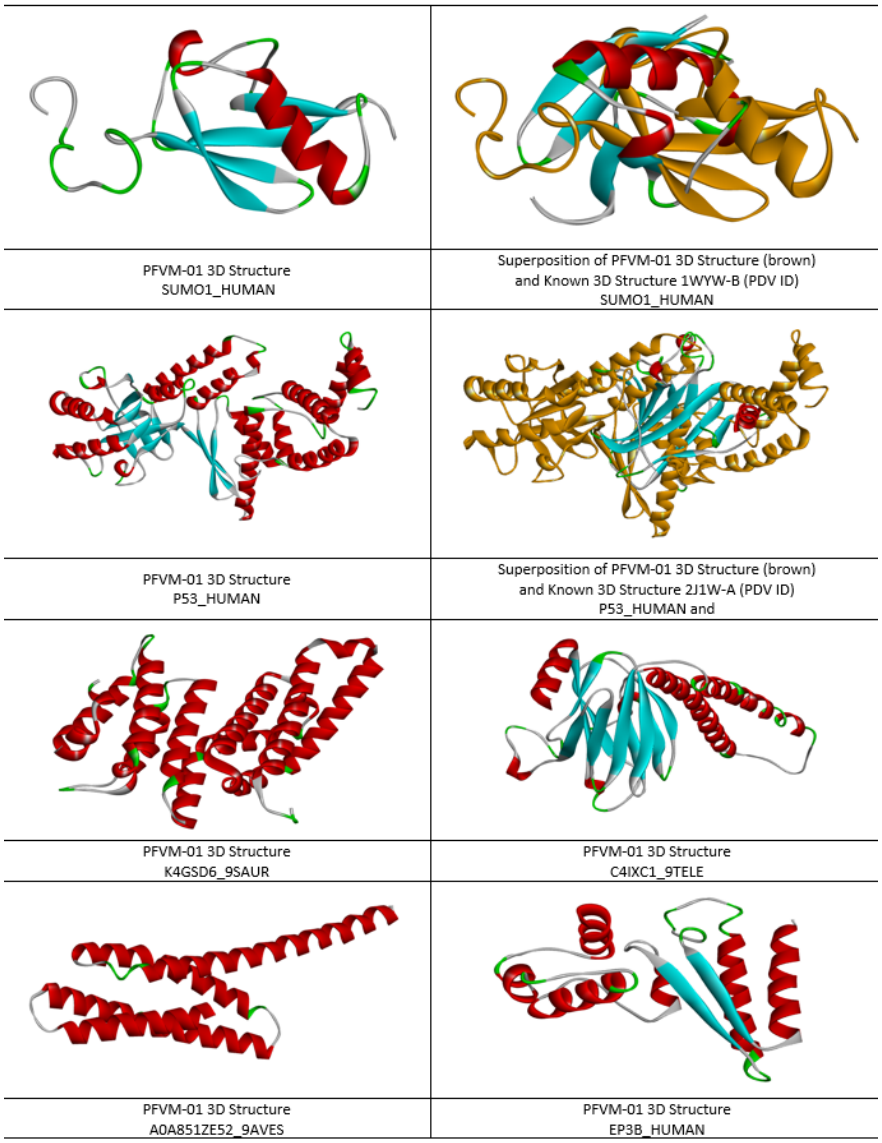


Figure 5. The predicted protein 3D structures from PFVM-01. First row displays the predicted 3D Structure from SUMO1_HUMAN PFVM-01; second row displays the predicted 3D Structure from P53_HUMAN PFVM-01; left side is comparison between known 3D structure and the predicted structure (brown color). The predicted 3D structures for K4GSD6_9SAUR, C4IXC1_9TELE, A0A851ZE52_9AVES and EP3B_-HUMAN are displayed respectively.

DISCUSSION

Computation and database for protein folding.

Many of computational methodologies and database for protein folding have been developed,11Compiani M, Capriotti E, "Computational and theoretical methods for protein folding". Biochemistry. 52 (48): 8601–24, (2013).. and the efforts may be divided into two aspects, one aspect is to predict the protein structure with thermodynamic stability and another aspect is to investigate the protein conformations with variability.

In first aspect, the prediction of protein structure from a sequence is pursuing to obtain a native folding

conformation with thermodynamic stability, and the stable structure is mainly controlled by hydrophobic interactions, hydrogen bonds, van der Waals forces, and conformational entropy. In general, the methods for prediction of protein structure fall into two main categories: template-free modeling and template-based modeling.22Guo JT, Ellrott K, Xu Y. A historical perspective of template-based protein structure prediction. Methods Mol Biol; 413:3–42, (2008).'33Dorn M, E Silva MB, Buriol LS, Lamb LC. Three-dimensional protein structure prediction: methods and computational strategies. Comput Biol Chem; 53PB:251–76, (2014).'44Brylinski M. Is the growth rate of Protein Data Bank sufficient to solve the protein structure prediction problem using template-based modeling? : Bio-Algorithms and Med-Systems[J]. Bio-Algorithms and Med-Systems, 11(1):1-7, (2015). The template-free methods, i.e., *ab initio* or *de novo*approaches, are based on the energy functions which carry out through the molecular dynamics (MD) simulation calculations under various force fields for atoms interaction or experiential parameters for group atoms interaction.55Honig B. Protein folding: from the levinthal paradox to structure prediction. J Mol Biol; 293:283–93, (1999).'66Onuchic JN, Wolynes PG. Theory of protein folding. Curr Opin Struct Biol; 14:70–5, (2004).'77Zhang J, Li W, Wang J, Qin M, Wu L, Yan Z, et al. Protein folding simulations: from coarse-grained model to all-atom model. IUBMB Life; 61:627–43, (2009). The protein with stable conformation is finally obtained by iterative convergence to lower thermodynamic free energy under defined force fields, such as AMBER,88Yang, L., Tan, C. H., Hsieh, M. J., Wang, J., Duan, Y., Cieplak, P., Caldwell, J., Kollman, P. A., and Luo, R. New-generation amber united-atom force field. J. Phys. Chem. B 110, 13166-13176, (2006). CHARMM99Brooks, B. etc, CHARMM: The biomolecular simulation program. J. Comput. Chem. 30, 1545-1614, (2009). and GROMOS1010Riniker, S., Christ, C. D., Hansen, H. S., Hunenberger, P. H., Oostenbrink, C., Steiner, D., and van Gunsteren, W. F. Calculation of relative free energies for ligand-protein binding, solvation, and conformational transitions using the GROMOS software. J. Phys. Chem. B 115, 13570-13577, (2011). force fields. The software from Chemistry at Harvard Macromolecular Mechanics (CHARMM) [33],1111Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". J Comp Chem. 4 (2): 187–217, (1983). is one of the most mature algorithm for molecular dynamics, which minimizes the free energy of a protein structure while collecting the molecular dynamics trajectory of united-atom all-atom, dihedral potential corrected variants and polarization. The Rosetta software1212Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popovic, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P., ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 487, 545-574, (2011). developed by the Berkeley Open Infrastructure for Network Computing Platform is one of *de novo* tools to predict protein structure, which is assembled by Monte Carlo simulated annealing procedure relying on a library of residue fragments.1313Kroese, D. P.; Brereton, T.; Taimre, T.; Botev, Z. I., "Why the Monte Carlo method is so important today". WIREs Comput Stat. 6: 386–392, (2014). In practice, the protein structure prediction is efficient for calculating smaller proteins, and requires vast computational resources for larger proteins. The template-based methods, such as homology modeling or comparative modeling, align sequences according to similarity of multiple templates from PDB, and then process energy optimization to predict protein 3D structure. With sequence homologous, it assumes that similar sequences have similar folding conformations. Depending on the homology modeling, I-TASSER1414Roy, A., Kucukural, A., and Zhang, Y., I-TASSER: A unified platform for automated protein structure and function prediction. Nat. Protoc. 5, 725-738, (2010)., Robetta1515Kim, D. E., Chivian, D., and Baker, D., Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 32, W526-W531, (2004). and MODELER1616Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Pieper, U., and Sali, A., Comparative protein structure modeling using MODELLER. Current Protocols in Protein Science, Chapter 2, Unit 2.9, (2007), Wiley, New York.'1717Liu, T., Tang, G. W., and Capriotti, E., Comparative Modeling: The state of the art and protein drug target structure prediction. Comb. Chem. High Throughput Screening 14, 532-537, (2011). software build protein for unknown 3D structure. If there is not a distinguishably similar sequence matched in PDB database, the template-free

approaches will provide the supplement for thermodynamics calculations. Recently, with a deep learning in artificial intelligence (AI), AlphaFold approach was particularly successful at predicting the most accurate structure and with demonstration in CASP13 and CASP14.1818DeepMind's protein-folding AI has solved a 50-year-old grand challenge of biology. MIT Technology Review. Retrieved (2020).,1919Sample, Ian (2 December 2018). "Google's DeepMind predicts 3D shapes of proteins". The Guardian. Retrieved 30 November (2020).,2020 "DeepMind's protein-folding AI has solved a 50-year-old grand challenge of biology". *MIT Technology Review* . (2020). AlphaFold first handled the protein structure as a spatial graph with the residues as nodes and the connection of residues as edges. Then, it trained the system on all available protein 3D structures from PDB together with the databases containing protein sequences of unknown structure. For physical interactions within proteins, it created an attention-based neural network system, and trained residue-to-residue and atom-atom using an internal confidence measure. The protein structure was refined by evolutionarily related multiple sequence alignment (MSA) and a representation of amino acid residue pairs. With iterating process, AlphaFold predicted the underlying physical structure of the protein and is able to determine highly-accurate structures.

In second aspect, the objective of protein folding is to investigate variations of conformations because the proteins in essence are non-static structures, but rather conformational ensembles with multiple states. With general knowledge, the protein adjusts the folding conformations under different environments or interaction of ligand or protein. Also, intrinsically disordered proteins and regions (IDPs/IDR) are widely distributed in natural proteins, which are associated with many biological processes and diseases.2121Chen J , Guo M , Wang X , et al., A comprehensive review and comparison of different computational methods for protein remote homology detection[J]. Briefings in Bioinformatics(2):2. 1–17, (2017). The IDPs/IDR for protein 3D structures can be identified by many experimental techniques.2222Robin van der Lee , etc. Classification of intrinsically disordered regions and proteins.[J]. Chemical Reviews, 114(13):6589, (2014). DisProt,2323Piovesan D, Tabaro F, Micetic I, et al. DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Res, 45:D1123–4, (2017). IDEAL2424Fukuchi S, Sakamoto S, Nobe Y, et al. IDEAL: intrinsically disordered proteins with extensive annotations and literature. Nucleic Acids Res; 40:D507–11, (2012). and MobiD2525Potenza E, Di Domenico T, Walsh I, et al. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. Nucleic Acids Res 2015;43:D315–20. are useful databases for IDP/IDR, and PDB also provides the illustration. Moreover, under physiological conditions, a native protein essentially is able to undergo a reversible transition between disorder and order folding conformations. In 1973, Anfinsen's Nobel prize-winning experiments2626Anfinsen CB, Principles that govern the folding of protein chains. Science 181: 223–230. showed that the protein ribonuclease can be reversibly denatured and re-natured in a test tube, and then over thousands of other proteins have been demonstrated with folding reversibility with condition changes. The protein has folding reversibility because of small energy barriers (5 to 15 kcal/mol) between the folded and unfolded populations.2727Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A (2006) Nucleic Acids Res 34:D204–D206, (1973). Different computational approach have been developed focusing on the variability of protein folding. In the late 1970s, Karplus and Weaver developed the diffusion-collision (DC) model,2828Karplus, M., and Weaver, D. L., Protein-folding dynamics. Nature 260, 404-406, (1976).,2929Karplus, M., and Weaver, D. L., Diffusion-collision model for protein folding. Biopolymers 18, 1421-1437, (1979).,3030Islam, S. A., Karplus, M., and Weaver, D. L., Application of the diffusion-collision model to the folding of three-helix bundle proteins. J. Mol. Biol. 318, 199-215, (2002).,3131Myers, J. K., and Oas, T. G., Preorganized secondary structure as an important determinant of fast protein folding. Nat. Struct. Biol. 8, 552-558, (2001). that explored the long-term protein evolution and allowed the large amplitude changes in the folding dynamics. Later it was modified into the foldon diffusion-collision (FDC)3232Fuxreiter, M., Simon, I., Friedrich, P., and Tompa, P., Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. J. Mol. Biol. 338, 1015-1026, (2004).,3333Compiani, M., Capriotti, E., and Casadio, R., Dynamics of the minimally frustrated helices determine the hierarchical folding of small helical proteins. Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys. 69, 051905, (2004).,3434Stizza, A., Capriotti, E., and Compiani, M., A minimal model of three-state folding dynamics of helical proteins. J. Phys. Chem. B 109, 4215-4226, (2005). which provided a more refined description of folding transforms, including predicting the secondary

native structure and specifying stability of the foldons themselves. In 1977, the hydrophobic collapse (HC) mechanism3535Dill, K. A., Theory for the folding and stability of globular proteins. Biochemistry 24, 1501-1509, (1985).,3636Haran, G., How, when and why proteins collapse: The relation to folding. Curr. Opin. Struct. Biol. 22, 14-20, (2012). was developed to predict that the hydrophobic forces and backbone forces result in chain collapse prior to the formation of elements of secondary structure. Of course, except hydrophobic, the hydrogen bonds and van der Waals forces are also steering the unfolded protein toward a collapsed configuration.3737Barbosa, M. A., Garcia, L. G., and Pereira de Araujo, A. F., Entropy reduction effect imposed by hydrogen bond formation on protein folding cooperativity: Evidence from a hydrophobic minimalist model. Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys. 72, 051903, (2005). In 2000, Folding@Home project was developed at Stanford University to compute the protein folding with widely adopting the contribution of computing resource. As a huge number of folding conformations, the molecular dynamics (MD) simulations is a time-demanding process which rely on parallel supercomputing architectures or using personal computing clusters.3838Zagrovic, B., Snow, C. D., Shirts, M. R., and Pande, V. S., Simulation of folding of a small α-helical protein in atomistic detail using worldwide-distributed computing. J. Mol. Biol. 323, 927-937, (2002).,3939Adcock, S. A., and McCammon, J. A., Molecular dynamics: survey of methods for simulating the activity of proteins.Chem. Rev. 106, 1589-1615, (2006).,4040Rizzuti, B., and Daggett, V., Using simulations to provide the framework for experimental protein folding studies. Arch. Biochem. Biophys. 531, 128-135, (2013).,4141Daggett, V., Protein folding-simulation. Chem. Rev. 106, 1898-1916, (2006). Anyway, the computational approaches for all possible conformations to thoroughly resolve the protein folding problem is now far less successful than was thought in the early days, and it is still one of challenging subjects in the field of protein physical science. Recently, Google's DeepMind applied the artificial intelligence (AI) and successfully developed Alphafold approach which can regularly predict protein structures with atomic accuracy competitive with experimental structures. It trained a neural network to accurately predict the distances between pairs of residues in a protein, and a protein was optimized by a simple gradient descent algorithm to realize structures. As the achievement of Alphafold, more scientific resource and attention are focusing on the resolution of protein folding problem.

The protein folding information can be extracted from protein structure databases. The PDB is the most inclusive repository of protein 3D structures. So far, nearly 190,000 protein 3D structures have been available in PDB, where approximately 90% are obtained by X-ray crystallography and the remain by NMR, CryoTEM and other techniques. The X-ray crystallography may determine accurate atomic coordination for 3D structure, but it only represents a specific static protein folding state. The NMR and CryoTEM display the protein flexibility that structural oscillation is limited around an equilibrium state under certain conditions. The Structure Classification of Proteins (SCOP)4242http://scop.mrc-lmb.cam.ac.uk/scop database classifies the protein structural domains into the hierarchy in terms of Species, Protein, Family, Superfamily, Fold and Class. It defines 1,232 folds, 2,026 superfamilies and 4,919 families. If two protein domains have similar secondary structures with the similar topological connections, they belong to the same fold. The Class, Architecture, Topological fold and Homologous superfamily (CATH) 4343http://www.cathdb.info classifies 95 million of protein domains into 1,391 topological folds and 6,119 superfamilies. If two proteins have similar topological fold and sequence in conjunction with similar functions, they are assumed to be associated with the same category in CATH. The ProTherm4444http://www.abren.net/protherm/ database is a source for understanding the protein folding stability with the thermodynamic parameters for 25,830 structures, which includes numerical data changes in Gibbs free energy, enthalpy, heat capacity and transition temperature etc. Nevertheless, the crucial question is whether the protein database can be directly utilized for the investigation of protein folding. The first question is if current protein structural data and future coming data are sufficient for fold recognition, and the answer is negative.[28] The second question is whether the defined topological folding patterns (about 1200 types of folds in SCOP and near 1,400 in CATH database) are enough to correlate the protein folding with the regulation of amino acid in sequence, and the answer is insufficient. However, a number of structural data from experimental and computational approaches should assist to understand the protein folding in some degree. As a whole, the longer fragments were hard thorough to investigate the folding patterns because of the larger the folding prototype involving less universal folding pattern. Therefore, to define a universal small folden as a prototype, such as the backbone of 5 amino acid

residues, may overcome these obstacles to probe the folding patterns in protein structure database.

Here, the protein structure fingerprint approach demonstrated a useful means to describe complete protein folding conformations and to construct explicit database for protein folding. In mathematical space, the backbone of 5 points connection is adopted as a universal folden and the complete folding space is described by 27 PFSC alphabetic letters. In biological space, the possible folds of 5 of amino acid residues are limited by constrains, and then different combinations of 5 of amino acid residues have different folding number and patterns. Thus, a database (5AAPFSC) was created to collect all folding shapes for all combinations of 5 of amino acid residues. For protein, one PFSC string represents a complete folding description, and one PFVM matrix represents comprehensive folding variation. Based on PFVM, not only does all possible folding conformations in astronomical number are obtained, but the most possible conformations are also obtained. Therefore, the protein structure fingerprint approach covers two aspects, it can predict stable folding conformation as well discover variations of folding conformation with massive number. Furthermore, the digital alphabetic PFSC provides a simplified mode to resolve the protein folding problem. As a result, the astronomical number of folding conformations can be easily stored into a database for protein folding. Thus, the protein structure fingerprint approach made a significant foundation to solve protein folding problem.

Image visualization vs. alphabetic description.

Due to complexity of protein structure, the protein structure fingerprint provided the PFSC alphabetical description to probe a huge number of protein data, especially it is suitable to study the protein folding conformations with an astronomical number. The protein 3D structure data are originally obtained by experimental measurements or computational approaches, which pursue to display 3D image visualization for protein structure. For single protein, its 3D structural image is displayed according thousand lines of atomic coordinates in the protein data file. Although a protein 3D structure is directly perceived through the senses to understand the folding orientation in space, it is not easily to illustrate the features of protein folding features. For comparison of proteins, with structural superposition, the similarity is quantified by the root-mean-square deviation (RMSD) as score. Nevertheless, it does not provide any detail where and how are similar or dissimilar between proteins, and artificial process severely affect the outcome. So, it is hard to explain the similarity and dissimilarity between proteins with 3D image visualization.4545Fitzkee NC, Fleming PJ, Gong H, Panasik N, Street TO, Rose GD. Are proteins made from a limited parts list? Trends Biochem Sci; 30:73–80, 2005.,4646Irving JA, Whisstock JC, Lesk AM. Protein structural alignments and functional genomics. Proteins; 42:378–382, 2001.,4747Sam V, Tai CH, Garnier J, Gibrat JF, Lee B, Munson PJ. ROC and confusion analysis of structure comparison methods identify the main causes of divergence from manual protein classification. BMC Bioinform; 7:206, 2006.,4848Yang J. Complete description of protein folding shapes for structural comparison[J]. Proteomics Research Journal, 3(1):1-22, (2012).,4949Sarah A. Middleton, Joseph Illuminati & Junhyong Kim, Complete fold annotation of the human proteome using a novel structural feature space, Scientific Reports volume 7, Article number: 46321 (2017). Furthermore, it is almost unimaginable to construct an astronomical number of 3D conformations for a protein to probe the protein folding problem, and to involve with billions of protein sequences even worse. However, one-dimensional PFSC alphabetic string provided a useful protocol to overcome these obstacles because it makes easily store and study a massive number of protein conformations. The PFSC alphabetic representation does not only simplify the description of protein conformation, but also it can align a large number of folding conformations for comparison. With advance, the PFSC alphabetic string covers the regular secondary fragments as well as the tertiary fragments, so it became a valuable approach to study the protein conformations with an astronomical number.

The alphabetic description has been adopted following development of protein structure study. Except to label regular secondary motifs of alpha helixes and beta strands, many different methods have been developed trying to label protein conformation more detail with alphabetic description. Some methods adopted more alphabetic letters to distinguish secondary structure motifs in detail which specified the patterns of hydrogen bonds and geometric criteria, such as Cα distances, Cα angles, dihedral angles between Cα atoms, or a pairs of ψ and φ dihedral angles around a Cα atom.5050Kabsch W, Sander C. Dictionary of protein secondary

structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers; 22:2577–2637, (1983).,5151Ridchards FM, Kundrot CE. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins; 3:71–84, (1988).,5252Frishman D, Argos P. Knowledge-based protein secondary structure. Proteins; 23:566–579, (1995).,5353Sklenar H, Etchebest C, Lavery R. Describing protein structure: a general algorithm yielding complete helicoidal parameters and aunique overall axis. Proteins; 6:46–60, (1989).,5454Labesse G, Colloc'h N, Pothier J, Mornon JP. P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. Comput Appl Biosci; 13:3:291–295, (1997).,5555Martin J, Letellier G, Marin A, Taly JF, de Brevern AG, Gibrat JF. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Struct Biol; 5:17–34, (2005). Other methods identified the patterns of structural segments with observations from a large number of structures in training database, and extracted certain motifs as folding prototypes by statistics adjustment and then labeled with alphabetic letters.5656Fetrow JS, Palumbo MJ, Berg G. Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. Proteins; 27:249–271, (1997).,5757Zhang X, Fetrow JS, Berg G. Design of an auto-associative neural network with hidden layer activations that were used to reclassify local protein structures. In: Crabb VJ, editor. Advances in Protein Chemistry. San Diego, CA: Academic Press; pp 397–404 (1994).,5858Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins; 41:271–287, (2000).,5959Alexandre G, de Brevern1, Valadie´ H, Hazout S, Etchebest C. Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. Prot Sci; 11:2871–2886, (2002).,6060Fourrier L, Benros C, Brevern AG. Use of a structural alphabet for analysis of short loops connecting repetitive structures. BMC Bioinform; 5:58, (2004).,6161Joseph A P, Srinivasan N, Brevern A G D. Improvement of protein structure comparison using a structural alphabet[J]. Biochimie, 93(9):1434, (2011). So far, most of alphabetic methods adopted 9-16 letters to describe various folding protocols with different lengths in fragments. Nevertheless, none of methods guarantee to provide a complete coverage for all possible folding patterns due to ignoring some of fragment motifs, such as irregular loops and coils or uncommon folding shape with rare appearances in structures, etc. However, the PFSC overcome the shortcomings, it provided a set of 27 alphabetical letters to cover all possible folds for successive 5 amino acid residues, and a PFSC string describe the complete folding conformation without gaps from N-terminus to C-terminus including regular secondary fragments and irregular tertiary fragments.

The protein structure fingerprint can describe the folding conformations with alphabetic description, no matter what the protein 3D structure is known or unknown. For protein with known 3D structure, the folding shape of each of 5 amino acid residues is assigned by one of PFSC letter according the atomic coordinates, and then the conformation of entire protein is expressed by a PFSC string. For protein without known 3D structure, the comprehensive folding variations for a protein are able simultaneously to be observed by the PFSC letters in PFVM with impressiveness covering all at one glance. Also, an astronomical number of folding conformations for a protein can be assembled with various PFSC letters in PFVM. Furthermore, any PFSC string represents one of folding conformations, and it can be conversely converted into 3D structure.

The alphabetic letters provide a brief description for biological structure in macromolecule system. The DNA polymer applies four letters (C, G, A and T) to describe the backbone strand comprised of four deoxyribonucleic acids in genetic code. The protein polymer applies 20 of amino acids with single letters to describe one-dimensional sequence. Biological structure is embedded in assembly processes, from one-dimensional DNA, mRNA to protein sequence until protein folding. In the first step, the genetic information is stored in the DNA sequence and transmitted through transcriptional and translated into one-dimension protein sequence. In the second step, the protein is folded from one-dimensional sequence to 3D structure for expressing the vitality of life. To date, however, the knowledge and understanding of protein folding lag far behind the DNA and protein sequences. The protein structure fingerprint made a significant progress which applied a set of 27 PFSC letters to describe protein folding. Thus, the PFSC perfectly matched alphabetic description of DNA and protein sequence, and it is possible to integrate the huge data of protein folding conformations with DNA or mRNA sequence and protein sequence.

Protein folding vs. the order of amino acids in sequence

It is well known that the protein folding in principal depends on the order of amino acids in sequence. Although researchers confirmed this principal with many biological experiments, it lacks a systematical depiction in bioinformatics aspect. Also, it is not easy to clearly illustrate how the order of amino acids in sequence affects the folding changes in protein. However, with a universal process, the PFVM integrally displays the correlation between protein folding changes and sequence variations. Generally, different protein sequences will have different folding patterns in PFVM. The folding pattern difference is presented in several aspects in PFVM even if only one amino acid was substituted. The differences include the changes of the types of folding shapes as well as the number of possible folding. Also, if one of amino acid is substituted, it will not only cause PFSC letter changes in one column, but a band of 5 columns in PFVM. These changes in PFVM well demonstrated that the protein folding depended on the order of amino acids in sequence.

The PFVM characteristically display the local folding variations along the sequence. The numbers changes of local folding shapes display the analogous fluctuation spectrum, and indicate some portions of protein with more flexible while other portions with less flexible. The fluctuation curves of numbers of local folding shapes for protein PDCD1_MOUSE and PDCD1_HUMAN are shown in Figure 6. First, each curve exposed how the folding flexibility following the order of amino acids in sequence. Second, both curves are different because of the differentiation of amino acids in sequences. At least, 4 locations in curves (35-43, 78-91, 139-151 and 170-179 in sequence) have the opposite tendency for the vibration of numbers of local folding variations. Thus, a fluctuation curve from PFVM concretely indicates how the protein folding relates the order of amino acid in sequence. Thus, the PFVM is a useful tool to probe the protein mutation, protein differentiation, protein design, protein prediction and protein misfolding etc.

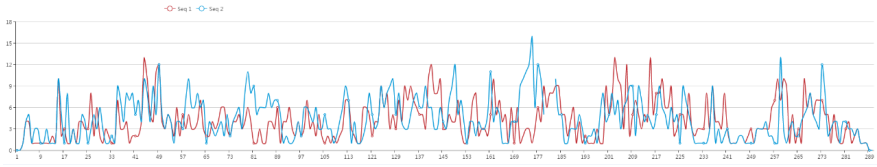


Figure 6. The fluctuation curves of numbers of local folding shapes for PDCD1_MOUSE (red) and PDCD1_-HUMAN (blue) proteins. The horizontal axis is the ruler of sequence, and vertical axis the number of local folding variations.

Disorder and flexibility in protein

The nature of protein conformation in biological system is not static, it is flexible following time being and condition change. The process involves intermediate folding states and stable folding states during their synthesis and degradation. These protein folding states may have various half-lifetime, and keep reversible disorder-order transitions. Also, the intrinsically disordered folds indeed exist in some of fragments or entire protein.11Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z, "Intrinsically disordered protein". Journal of Molecular Graphics & Modelling. 19 (1): 26–59, (2001).·22Dyson HJ, Wright PE, "Intrinsically unstructured proteins and their functions". Nature Reviews. Molecular Cell Biology. 6 (3): 197–208, (2005)..·33Dunker AK, Silman I, Uversky VN, Sussman JL, "Function and structure of inherently disordered proteins". Current Opinion in Structural Biology. 18 (6): 756–64, (2008).

The results from experimental measurement are actually affected by intrinsically disorder and flexibility in protein. About two-thirds of data in PDB do not have the complete 3D structures covering entire sequence by X-ray crystallography measurement because of the unobserved regions that frequently correspond to the disorder.44Le Gall T, Romero PR, Cortese MS, Uversky VN,Dunker AK. Intrinsic disorder in the Protein Data Bank. J Biomol Struct Dyn; 24:325-42, (2007). With nuclear magnetic resonance (NMR) spectroscopy,

it can reveal an ensemble of protein structures with flexible dynamics or disorder states limited around a static state. With transmission electron cryomicroscopy (CryoTEM), it can provide new insights for protein structure with large assemblies which allow observation of protein structures in their native environment at cryogenic temperature. With computational simulations, it can obtain protein dynamics trajectories which reveal the disorder and flexibility within protein structure. The protein folding occur in many steps, and it may spend nearly 96% time in some states, and also in various intermediate states with minimum thermodynamic free energies in energy landscape.55Heath Ecroyd; John A. Carver, "Unraveling the mysteries of protein folding and misfolding". IUBMB Life (review). 60 (12): 769–774, (2008). ,66Robert B Best, "Atomistic molecular simulations of protein folding". Current Opinion in Structural Biology (review). 22 (1): 52–61, (2012).

Under physiological condition, the protein flexibility substantially plays important role for biological functions. Even after self-assemble into a native state with active function, some parts of proteins may remain folding variations.77Berg JM, Tymoczko JL, Stryer L, "3. Protein Structure and Function". Biochemistry. San Francisco: W. H. Freeman. ISBN 0-7167-4684-0, (2002). The protein biological functions associate with either the native stable structure or the dynamics motion in structure. It is undeniable that the protein structures with both dynamic disorder conformations and stable conformations subsequently are linked to important functions such as allosteric regulation and enzyme catalysis.88Bu Z, Callaway DJ, "Proteins move! Protein dynamics and long-range allostery in cell signaling". Advances in Protein Chemistry and Structural Biology. Advances in Protein Chemistry and Structural Biology. 83: 163–221, (2011).·99Kamerlin SC, Warshel A, "At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis?". Proteins. 78 (6): 1339–75, (2010). Although the so-called native conformation is essential to a specific function in cell, the environmental factors, ligands and other proteins may cause the native conformation into altered folds which may trigger to act different biological function, such as active or inhibiting function or toxic affection. This is actual process when a drug targets proteins to make the conformation changes and to cause activation or inhibition for protein function. Overall, a complete ensemble of conformations really represents the nature of protein folding for multiple functions in physiological condition or various environments. It is significant that the PFVM provides the comprehensive local folding variations for discovery of various conformations in astronomical number. Of course, the remaining question is how to distinguish local folding variations in PFVM with association of function changes. Ultimately, we need better experimental data and further data mining to understand the role of each local folding variation in biological functions.

CONCLUSION

The protein structure fingerprint provided a significant means to probe the protein folding problem, especially overcame the obstacles how to reveal and handle an astronomical number of folding conformations. The folding shapes of 5 amino acid residues, as a universal element, are comprehended, and then expanded over sequence to expose an astronomical number of folding conformations. A set of 27 PFSC alphabetical letters are represented as digital expression. The local folding variations in PFVM really uncovered how the protein folds are correlated by the order of amino acids in sequence. Also, it provides the prospect to construct all possible conformations with astronomical number as well as to predict the most probable conformations for a protein. These advantages information may promote the research in protein structures, such as the protein folding, protein structure prediction, intrinsically disordered protein, protein mutation and protein design etc.

DATA AVAILABILITY

- All sequences are from UniProt database: https://www.uniprot.org/
- All protein 3D structures are from PDB: https://www.rcsb.org/
- The PFSC and PFVM can obtained from website http://www.micropht.com/ or http://pfsc.considerin.net

COMPETING INTERESTS STATEMENT

We declare that we have no competing interests.

ACKNOWLEDGEMENTS

The protein structure fingerprint scheme has been developed by author Jiaan Yang and published on "Yang J, Comprehensive description of protein structures using protein folding shape code. Proteins 2008;71.3:1497-1518"

Supplementary Information:

The supplementary information was stored in a file with rich text format (RTF), which hold longer sequence in same row and better display the PFVM matrix. The name of supplementary file name is "Supplementary-PFVM-Matrix-for 14 proteins.rtf".

Author Contributions

G.W. involved input data preparation; W.X.C. involved output data collection and analysis; X.F.Z. wrote the software code; Q.H, H.G, Q.Q, X.J and L.Z. verified algebraic calculations; S.T.S. deployed software code to web page; P.Z. tested and verified the code and final data; J.Y. overall designed and managed the project and made the manuscript preparation.

Contact Information

Correspondence and requests for materials should be addressed to J.Y. (jyang@micropht.com)

REFERENCE