NOVA
SCIENCE PUBLISHERS, INC.

COMPLETE DESCRIPTION OF PROTEIN FOLDING SHAPES FOR STRUCTURAL COMPARISON

JIAAN YANG

400 Oser Avenue, Suite 1600
Hauppauge, N. Y. 11788-3619
Phone (631) 231-7269
Fax (631) 231-8175
E-mail: main@novapublishers.com
www.novapublishers.com

ISBN: 978-1-61122-595-2 2011

MATHEMATICS RESEARCH DEVELOPMENTS

COMPLETE DESCRIPTION OF PROTEIN FOLDING SHAPES FOR STRUCTURAL COMPARISON

MATHEMATICS RESEARCH DEVELOPMENTS

MATHEMATICS RESEARCH DEVELOPMENTS

Additional E-books in this series can be found on Nova's website under the E-book tab.

MATHEMATICS RESEARCH DEVELOPMENTS

COMPLETE DESCRIPTION OF PROTEIN FOLDING SHAPES FOR STRUCTURAL COMPARISON

JIAAN YANG

Nova Science Publishers, Inc.
New York

For permission to use material from this book please contact us:
Telephone 631-231-7269; Fax 631-231-8175
Web Site: http://www.novapublishers.com

NOTICE TO THE READER

The Publisher has taken reasonable care in the preparation of this book, but makes no expressed or implied warranty of any kind and assumes no responsibility for any errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of information contained in this book. The Publisher shall not be liable for any special, consequential, or exemplary damages resulting, in whole or in part, from the readers' use of, or reliance upon, this material. Any parts of this book based on government reports are so indicated and copyright is claimed for those parts to the extent applicable to compilations of such works.

Independent verification should be sought for any data, advice or recommendations contained in this book. In addition, no responsibility is assumed by the publisher for any injury and/or damage to persons or property arising from any methods, products, instructions, ideas or otherwise contained in this publication.

This publication is designed to provide accurate and authoritative information with regard to the subject matter covered herein. It is sold with the clear understanding that the Publisher is not engaged in rendering legal or any other professional services. If legal or any other expert assistance is required, the services of a competent person should be sought. FROM A DECLARATION OF PARTICIPANTS JOINTLY ADOPTED BY A COMMITTEE OF THE AMERICAN BAR ASSOCIATION AND A COMMITTEE OF PUBLISHERS.

Additional color graphics may be available in the e-book version of this book.

LIBRARY OF CONGRESS CATALOGING-IN-PUBLICATION DATA

ISBN: 978-1-61122-595-2

Published by Nova Science Publishers, Inc. † New York

CONTENTS

PREFACE

The effort to develop better description of protein three-dimensional folding structures has dominated biochemistry and drug discovery research for more than 70 years since Pauling first defined the helical configurations as secondary structure for protein in 1940. The challenge is how to acquire a complete description of protein folding shapes from N-terminal to C-terminal, including regular secondary structure as well as irregular tertiary structure. Here, a novel description method is introduced, which a set of 27 vectors is rigorously derived mathematically from an enclosed space. Each vector represents a three-dimensional folding shape of five successive C_α atoms, and the protein conformation can be completely described along protein backbone. These vectors are expressed by 27 alphabetic symbols, which are called as protein folding shape code (PFSC). Consequently, with PFSC, the folding conformation of any protein with given three-dimensional structure is able to be converted into a simple one-dimensional alphabetic string without gap. Furthermore, to take the advantage of one-dimensional description of folding shapes, the protein conformational structures are able to be compared with Needleman-Wunsch alignment algorithm. The global similarity of protein 3D structures is able to be assessed by a value of protein folding structure alignment score (PFSA-S) as a quantitative measurement, and the similarity and dissimilarity of local structures is able to be examined by alignment table. The results show that this approach has the capability not only to distinguish protein conformers with relatively high similarity, but also to compare proteins with diverse degrees in structural homology. Therefore, this approach provides a consistent procedure, and it produces a unique score for assessment of similarity in protein structure comparison. The significant is that the complete description of protein folding shapes provides a simple and effective means to screen protein database, compare protein structures, search protein fragment and probe drug binding site, study protein mutation and protein misfolding and so on.

Chapter 1

INTRODUCTION

Historically, after Pauling and Corey discovered the helical structure of protein in 1940, scientists recognized that amino acid polypeptide chain may fold into regular secondary structures which stabilized by peptide hydrogen bonds [1-2]. In 1952, the concept of three levels of protein structure is introduced by Linderstrom-Lang as primary, secondary and tertiary structures [3]. The primary structure is defined by protein sequence which is the link of polypeptide of amino acid and does not describe the spatial arrangement. The secondary structure is the regular three-dimensional form of local segments of proteins, such as α-helix and β-strand, which are determined by patterns of hydrogen bonds between backbone amide and carboxyl groups. The tertiary structure refers to the spatial arrangement of entire protein, which is frequently stabilized by the sequestration of hydrophobic amino acid residues in the protein core. In tertiary structure, the secondary structural fragments are linked by coils and loops, and are formed into folding units at higher level in hierarchy of protein [4-5]. In 1985, the quaternary structure is defines by Bernal as the fourth level of structure, which describes the arrangement of multiple folded protein molecules in complex [6].

Protein 3D structures are determined by either X-ray crystallography or nuclear magnetic resonance (NMR), and are deposited at Protein Data Bank (PDB) [7]. Except experimental measurement, a number of methods for computational prediction of protein structure has been developed, which use the sequence of amino acid, existing protein structures and homology modeling to build 3D model for a protein [8-

15]. With accurate structural coordinates at atomic level for a protein, the 3D structure is able to be presented as the protein modeling for visualization. However, it is so difficult to express the structure of protein clearly and compare 3D structures directly. Although, more and more protein structures are determined, it is hard mining protein structures through the mass of computer database for the application of quantitative structure-activity relationship (QSAR). The development of a complete description for protein folding conformation is significant step to overcome these difficulties. Here, a novel description method, protein folding shape code (PFSC), is introduced. The applications of PFSC demonstrate the effect in protein comparison and the capabilities screening protein database.

Chapter 2

STRUCTURAL CHARACTERISTICS OF PROTEIN CONFORMATION

In protein, the fold of the polypeptide backbone surrounding each C_α carbon can be illustrated by two torsion angles: phi (Φ), along the N-C_α bond, and psi (Ψ), along the C_α-C bond at each amino acid residue [2, 16-18]. However, the torsion angles for most amino acids are restricted and appeared as clusters in the Ramachandran plot [19], while the torsion angles for amino acids of proline and glycine are beyond the common ranges of other residues. Therefore, the distribution of torsion angles Φ and Ψ indicates both of restriction and arbitrariness because of complexity of protein structures.

A protein 3D structure is a composite of various fragments as local structures. The regular secondary structure, α-helices and β-strands, is local fragments with repetitive patterns along the protein backbone as a result of hydrogen bonding [9-12]. The β-turns and reverse turns are two other types of local structures involving a change of the overall chain directions [20-21]. Statistically, these three types of local fragments only occupy about 50-55% of the overall protein structures [22]. The remaining local structures are irregular coils or loops that are difficult to be identified and described. To date, various methods have been developed to attempt overcoming the obstruction in description of overall protein structure.

Chapter 3

METHODS FOR DESCRIPTION OF PROTEIN CONFORMATION

PROTEIN STRUCTURAL CLASSIFICATION

The development of protein structural classification provides a visual recognition of recurrent folding patterns. The structural classification is based on topological arrangements of protein secondary structures [23]. Three well known databases, SCOP, CATH and FSSP, have been created to store protein structural classification information [24-25]. Working with these databases, many algorithms have been developed to compare proteins with alignment of secondary structures and 3D coordinates of C_α atoms in structures, such as DALI [26], STRUCTAL [27], VAST [28], LOCK [29], 3DSearch [30], CE [31], SSM [32] and PALI [33]. Although about 10,000 protein structural classification families are predicted to exist in all organisms [34], at high hierarchy of classification, all of known protein structures are sorted into four broad structural classes: all-α, all-β, α/β and α+β. These four classes differ on secondary structure compositions and β-sheet topologies, which contain certain global characteristics of protein folding shapes [35]. The structural classification is one step toward comprehensive understanding protein folding.

VARIOUS METHODS FOR DESCRIPTIONS OF PROTEIN FOLDING STRUCTURES

In spite of the increased knowledge of protein structures that has allowed more understanding the relationship between sequences, structures, functions and properties, the description of protein folding shape is still a challenging subject [36-37]. A number of approaches have developed to improve the descriptions for protein folding structures.

SDP [38]. The Sequence-Derived Prediction (SDP) method assigns probe amino acid sequences to known 3D protein structures in order to predict protein folds. The SDP method defines new functions that combined amino acid-to-structure compatibility scores with sequence-derived properties-to-structure compatibility functions, which improves the assignment of sequences to known 3D folds.

DSSP [39]. The Database of Secondary Structure of Proteins (DSSP) program is a widely accepted method for assigning secondary structures to proteins. Using a pattern-recognition process, hydrogen-bonded and geometrical features are extracted from protein 3D coordinates. The cooperative secondary structure was recognized as repeats of the elementary hydrogen-bonding patterns of α-helices and β-strands. The result is a compilation of primary structure and secondary structure of globular proteins. The secondary structural assignments for most proteins in PDB are generated by DSSP.

DEFINE [40].The Define Structure (DEFINE) method is a computer program that describes secondary structures of proteins by using difference distance matrices of C_α atoms, and then supersecondary structures are obtained by using the secondary structures as straight lines segments. This program provides accurate two-dimensional display of the 3D structure.

STRIDE [41].The Structural Identification (STRIDE) method is a software tool which assigns protein secondary structures from atomic coordinates based on combination of hydrogen bond energy and statistical derived Φ and Ψ torsion angle information around C_α atoms. Results are optimized with verified set of secondary structural elements in PDB. Assignment of secondary structures by STRIDE shows good agreement with DSSP.

PCURVE [42]. P-Curve (PCURVE) is a computer program that obtains helicoidal structures from the atomic coordinates of peptide backbone, and yields helical axis representing the overall folding of protein. Each portion of the axis generated by P-Curve is reliant on a minimum of 9 peptide units. This method generates a set of 16 parameters for analyzing, comparing, or reconstructing protein backbone geometry.

PSEA [43]. Protein Secondary Element Assignment (PSEA) is a method that assigns secondary structural elements relied solely on the protein C_α coordinates. The parameters of angle, dihedral angle and distances between C_α atoms are used to perform the task as efficiently as other methods based on backbone analysis. In overall, the PSEA assignment is agreement with various other methods.

KAKSI [44]. KAKSI is a method for secondary structure assignment that is based on a set of geometrical values of C_α distances and Φ and Ψ dihedral angles around C_α atoms. The parameters of KAKSI are chosen with best fitting α-helices and β-sheets extracted from the PDB data. This method focuses on improvement of the termini of secondary structural segment with appropriate length.

SBB [45]. Structural Building Blocks (SBB) method is an extension of autoassociative artificial neural network (autoANN) [46] with the classification of seven residue protein segments. 6 common occurring patterns of SBB are defined from the statistical analysis with using database of 116 different protein chains. The SBB method identifies the regular secondary structures with the caps at ends for helices and strands, and distinct patterns of SBB occur in the random coil regions of proteins.

PB [47-49]. Protein Block (PB) is a method that identifies various folding patterns of five consecutive C_α atoms from 342 proteins. The 16 of protein blocks are selected from 86,628 folding patterns according a suitable balance between a correct approximation of 3D structures with an average RMSDA (root mean square deviation on angular values) of

30 degrees and acceptable initial prediction accuracy. The selected protein blocks are represented by 16 alphabetic letters.

Approaches for protein folding structural descriptions can be divided into three categories according to the strategy and parameters employed. The first category is protein homology methods that predict the secondary structures from given sequence of amino acids. The second category includes methods that primarily manipulate various geometric parameters based on protein 3D structural information. Many methods assign the regular secondary structure elements and super secondary motifs based on geometric criteria, such as C_α distances, C_α angles, dihedral angles between C_α atoms, pairs of Φ and Ψ dihedral angles around a single C_α atom, and specific patterns of hydrogen bonds. The third category is the local structural methods that identify the patterns of structural segments with observations from large number of structures in protein library and database, and then define certain motifs as folding prototypes with statistics adjustment. In general, most of methods show a broad agreement for α-helices and β-strands assignment. Various methods offer different lengths for the secondary structural segments because of the assignment difference at the beginning and end of segment.

CHALLENGE OF DESCRIPTION OF PROTEIN CONFORMATION

It is difficult to properly describe the complicated folding protein structures [50]. Three aspects challenge the description of protein folding structures. The first consideration is how to describe the irregular loops and coils, which take place about 40% in overall protein structures with varieties of folding patterns. It has been estimated about 4000 possible types of folding in protein structures, among which about 2,000 types were known in naturally-occurring proteins [51]. So, about 2000 of remaining folding types are unknown or rare. An ideal protein folding shape description is expected to be able to cover the regular secondary and irregular tertiary structural fragments as well as to reserve all possible folding patterns for shapes with common and rare appearances in structuirres. The second consideration is how to describe the protein folding with higher accuracy using the limited number of the selected folding patterns. It is particularly difficult to distinguish minor changes within structural fragments regardless of whether they are regular secondary structural elements or irregular turns. The third consideration is how to illustrate the foding descriptions with explicit structural meaning along proteon backbone. Therefore, a good protein description method is expected to have capabilities to cover all possible folding shape patterns, to reveal similarity and dissimilarity with proper sensitivity for protein comparisons and to provide meaningful explanation for protein folding descriptions.

Chapter 4

COMPLETE DESCRIPTION OF PROTEIN FOLDING SHAPES

A novel approach, protein folding shape code (PFSC) [52], provides a complete description for protein 3D conformation. With PFSC, a set of 27 PFSC vectors in Figure 1 is mathematically derived from an enclosed space to represent all possible prototypes of folding shapes for each five successive C_α atoms. The 27 PFSC vectors are able to map all possible folding shapes, including the regular secondary structure and irregular tertiary structure. In other words, the 27 PFSC vectors are capable completely to describe the change of protein folding shapes along protein C_α atom backbone from N-terminus to C-terminus without gap.

Three blocks in Figure 1 represent three regions of pitch distance of five successive C_α atoms, and the nine vectors in each block represent the nine folding shape patterns determined by three regions from each of two torsion angles; each vector is represented by a letter, a folding shape pattern and an arrow. The 27 vectors are then symbolized by the 26 alphabetic letters and the symbol “$”. Therefore, each letter represents a specific prototype of folding shape pattern of five successive residues. For example, the letter “A” represents a typical α-helix and “B” a typical β-strand. Other letters represent various possible folding shapes, including shapes being partial analogous to α-helix, β-strand, loop or coil and so on. Next, the vector characteristic is represented by an arrow line, which the initial and terminal points represent the N- and C-termini for the PFSC vector respectively. The “α”, “β” or “*”at each end of vector indicates the folding features similar to α-helix, β-strand or random coil respectively.

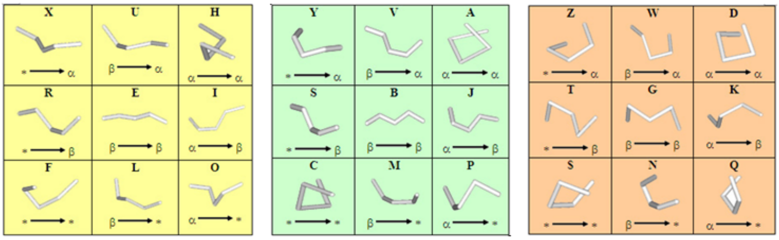


Figure 1. Diagram of 27 protein folding shape code (PFSC) vectors. Three blocks represent three regions of pitch distance; the nine vectors in each block represent the folding shape patterns determined by two torsion angles; each vector is simultaneously represented by a letter, a folding shape pattern and an arrow. The vector characteristic is represented by an arrow line, which the initial and terminal points represent the N- and C-termini for the PFSC vector respectively. The "α", "β" or "*"at each end of vector indicates the folding features similar to α-helix, β-strand or random coil respectively.

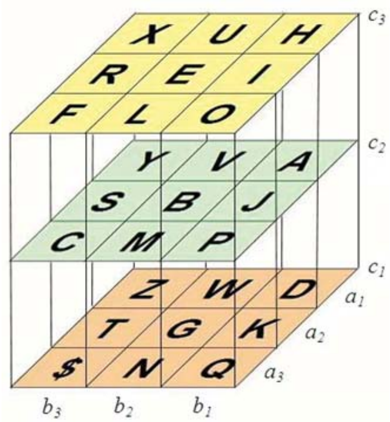


Figure 2. Relationship of PFSC 27 vectors. Three layers represent different pitch distance blocks, and the nine vectors in each horizontal layer are the results of combination changing from two torsion angles. The nine vectors in each vertical slice belong to same zone of a torsion angle. Here a_i is the index of the first torsion angle and b_j is the index of the second, for five successive C_α atoms; c_k is the index of pitch distance between two terminal atoms.

The three-dimensional arrangement of 27 vectors in Figure 2 shows the integral relationship of PFSC. Three axes of *a*, *b* and *c* represent three components, i.e. two torsion angles and one pitch distance. Each component is partitioned into three ranges, which creates the 27 PFSC vectors. Each vector associates to other vectors in horizontal and vertical directions. Also, a vector shares certain folding features with the surrounding vectors, and the vectors in horizontal layer or the vertical slice share a common feature. Furthermore, the 27 vectors are not the isolated folding patterns, and they can be transferred each other in an enclosed space.

Any protein conformation is able to be completely described along backbone by one-dimensional folding shape description with 27 PFSC vectors. Figure 3 explains how to obtain one-dimensional folding shape description from N-terminal to N-terminal with PFSC vectors of five successive C_α atoms. Here typical α-helix is remarked by red color, typical β-strand blue color, folding shape analogous to α-helix or –strand pink color and irregular folding shape black color.

Chapter 5

PROTEIN STRUCTURAL COMPARISON

SIMILARITY SCORE AND ALIGNMENT TABLE

With one-dimensional folding shape description, the protein conformation structures are compared by protein folding shape alignment (PFSA) approach [53]. Similar to the protein sequence alignment, one-dimensional alphabetic strings for protein folding shape descriptions are compared by an alignment algorithm. The Needleman-Wunsch algorithm of dynamic programming technique [54] is used in the PFSA structural alignment. Therefore, the structural similarity of two proteins is able to be discovered by an optimized alignment.

Based on the optimized alignment, the protein structural similarity score is calculated. Each match of identical folding shape is assigned by 2; analogous folding shape 1; dislike folding shape 0; penalty of open a gap -2 and penalty of extended a gap -0.5. The value of protein folding structure alignment score (PFSA-S) is determined by the total contribution of identical folding shapes, analogous folding shapes and gaps [54].

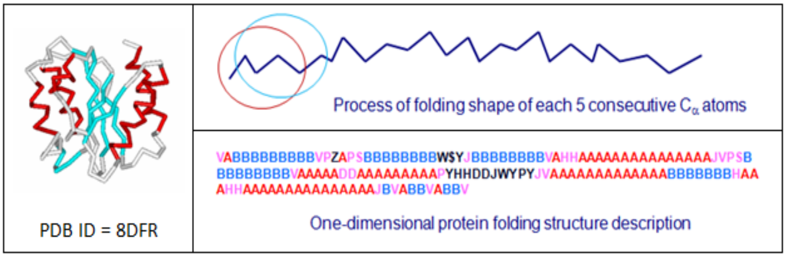

PDB ID = 8DFR
Process of folding shape of each 5 consecutive C_α atoms
VABBBBBBBBBVPZAPSBBBBBBBBW$YJBBBBBBBBVAHHAAAAAAAAAAAAAAAJVPSB
BBBBBBBBVAAAAADDAAAAAAAAAPYHHDDJWYPYJVAAAAAAAAAAAAABBBBBBBHAA
AHHAAAAAAAAAAAAAAAJBVABBVABBV
One-dimensional protein folding structure description

Figure 3. One-dimensional folding shape description. The left panel display the 3D structure of protein (PDB ID = 8DFR); top right panel shows how to apply a vector of five successive C_α atoms; bottom right panel the one-dimensional folding shape description.

The consequence of alignment of one-dimensional alphabetic strings for protein folding conformations is to obtain the PFSA alignment table for structural comparison. There are two types of alignment tables, i.e. sequence-dependence mode and sequence-independence mode. For conformational analysis of same protein or proteins with mutation, the structural alignment may prefer the sequence-dependent mode as the insertion of gap is not necessary. For different proteins with unlike sequences and size, the structural alignment takes the advantage of the sequence-independent mode, which allows inserting gaps in order to obtain the best match for local structure. The PFSA alignment table has several features. First, the alignment table is able to reveal the similarity and dissimilarity explicitly for local structure. Second, the alignment table exhibits how all similar fragments are matched or shifted with insertion of gaps. Third, in the alignment table the structural folding shape associates with the corresponding residue of five consecutive amino acids, which shows the transform of folding shape following the change of residues.

Conformation Analysis with Sequence-Dependent Mode

As the conformers for same protein generally have similar structures, the comparison requires the approach with higher sensitivity to distinguish structural alike conformation. The dynamic structures of protein in solvent can be measured by NMR techniques, and the certain number of conformations is collected and required to be analyzed for structural characteristics. For conformational structures with higher similarity, it is hard to distinguish each other [55]. For example, 30 conformers of amyloid β-peptides (Aβ-42) from 1Z0Q (PDB ID) are superimposed in Figure 4, and it is not easily to discover the structural similarity and dissimilarity. However, the PFSA provides the conformational analysis, which is displayed in alignment table in Table 1. The complete description of protein folding shape is able explicitly to illustrate the structural characteristics in alignment table. The alignment table provides better analysis for conformations than structural superimposition, which clearly reveals the fragment of residue 9-22 with structural stability and the fragment of residue 23-42 with flexible structure at N-terminal. Also, it reveals that the unstable fragment of residue 23-42 has dynamic structural characteristics of β-stand, which location is overall agreement with the protein misfolding involving with alzheimer disease [56].

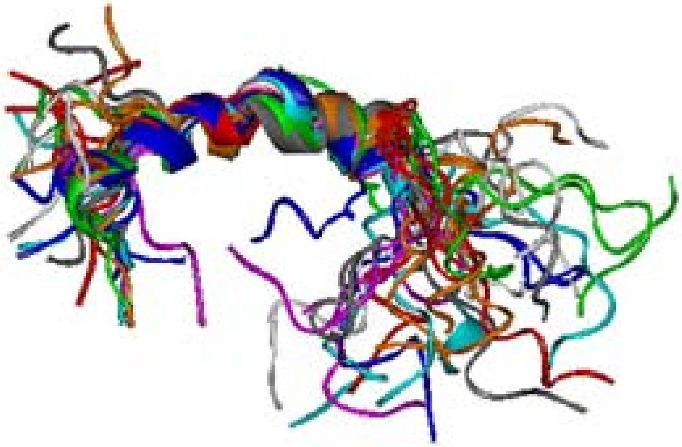


Figure 4. 30 conformers of amyloid β-peptides (Aβ-42) from 1Z0Q (PDB ID).

Table 1. Alignment Table for Comparison of 30 conformers of 1Z0Q (PDB ID)

Ruler	1 2 3 4
	123456789012345678901234567890123456789012
Seq.	daefrhdsgyevhhqklvffaedvgsnkgaiiglmvggvvia
1z0q01	PZAAAADAAAAAAPZAADAAAQZAAAAAAAAAJWYAJW
1z0q02	BHAAAAAAAAAAAQZAADAADAAJWZAAQSWZJWZABB
1z0q03	AAAQZAAAAAAAAPZAADAAAAAPC$CSBVAAAAAADJ
1z0q04	AAAAHAAAAAAAAAAAADAAAADAPYAAAAAAAQZPSW
1z0q05	PZAAHAAAAAAAAPZAAAAAAQZAQSWSVAAAAQZJBV
1z0q06	VJVQYAAAAAAAAPZAADAADDAJVAAAAAPZJWZABB
1z0q07	BVADAADAAAAAAPZAAAAAAQCSWZAAAHHAAAADJW
1z0q08	VAAAJVQYAAAAAQZAADAADDDJVAAAAAAAAQZPSB
1z0q09	CZAAAAQYAAAAAPZAAAAADDAJWZAAABVAJWZJBV
1z0q10	HHAJBVAAAAAAAPZAAAAAADAAAAAAHAAAAAAADJ
1z0q11	CZAAAAQYAAAAAPZAAAAAAQZADAQSHHADAAQSAP
1z0q12	AAQZAAAAAAAAAAAAADAAAQZAPZAAAAAAAQZJWY
1z0q13	HHAAJVAAAAAAAPZAAAAAAADJVAAAAAAAJWGYHH
1z0q14	WYAAAADAAAAAAPZAADAADDAJWSWSVAAAAQZPYH
1z0q15	VAADAAAAAAAAAAAAADAAAQZAAAAAJBWCSVADJW
1z0q16	AHAAAAQZAAAAAPZAAAAAAAAAWZAAWSAAAAQCYH
1z0q17	VAAAAADAAAAAAPZAAAAADDAJVQ$SVAAAAAAQSW
1z0q18	PZAQZAAAAAAAAQZAADAAAADJW$CSVJVAJWYAAH
1z0q19	AAAAAAQYAAAAAPZAAAAAAAAAJVAAAHAAAWZPYA
1z0q20	VAAJVADAAAAAAAAAAAAADDDAPZAAAAAAAQZPSW
1z0q21	WZAAAAAAAAAAAAAAAAAAAQCYJVQSBVAAAQZPSV
1z0q22	JVAAHAAAAAAAAAAAAAAAADDAPZAAAAAAJWZAHB
1z0q23	PZAAAAAAAAAAAPZAADAAAAAPSAAAAAADAHPSBA
1z0q24	AJVJVAAAAAAAAPZAADAAAAAJVAAAJBWYJVAAAA
1z0q25	AJVADAAAAAAAAAAAAAAAAADAPZAAAAAAJWYAJW
1z0q26	HHADAAAAAAAAAAAAADAAAAAQZAAAAAADJWZAJV
1z0q27	HAAAAAAAAAAAAPZAADAAAADJBVAABBAAJWZAJB
1z0q28	AAAAAAQZAAAAAPZAAAAAAAAJW$CSVJVAAQZPSW
1z0q29	HAAAAADAAAAAAAAAADAAAAAAQSWSVJVDJWZAHA
1z0q30	BVAQZAQZAAAAAAAAAAAAAQCZJVAAQYAAAQZPSW

The names of conformers are listed on left column. The amino acid sequence and rule for number of residue are listed on top rows. The protein folding shape code (PFSC) for each conformer is listed following the structure name. The typical α-helices are remarked with red color, the β-strands with blue color and the tertian fragments with black. Also, these analogue structures with secondary structure are remarked with pink color.

The conformational structures for same protein may be generated for various projects and different circumstances, and they are measured by X-ray crystallography or NMR. For example, images of 12 structures of protein enzyme of Dihydrofolate reductase (DHFR) from PDB database are displayed in Figure 5. Each enzyme has same sequence with 159 amino acids, and their structures are acquired by different conditions and laboratories. Also, these protein complexes may contain different number of ligands, such as methotrexate (MTX) and nadp nicotinamide-adenine-dinucleotide phosphate (NAP) etc. For example, to compare 1RA1-A with other structures, the variation of structural conformations is obviously revealed by alignment table in Table 2 while the global protein similarity is quantitatively assessed by the protein folding structure alignment score (PFSA-S) in Table 3.

Comparison of Proteins with Sequence-Independent Mode

Different proteins not only have various sequences of amino acid, but also the lower similarity in structure, such as different folding pattern, different topologic distribution and different size. These factors increase the difficulty to compare protein structures in 3D geometric space. Since the protein folding shape code (PFSC) is able completely describe the folding conformation of protein along backbone, the comparison of protein folding conformation is easily to be displayed by alignment table. Figure 6(A) and (B) display the 3D structural images of protein granulocyte-colony stimulating factor (G-CSF) of 1BGE-B (PDB ID) and protein granulocyte-macrophage colony-stimulating factor (GM-CSF) of 2GMF-A (PDB ID), which belong two different families respectively, i.e. long-chain cytokines and short-chain cytokines, with different length of sequences. Figure 6(C) shows one of superimpositions of 1BGE-B and of 2GMF-A with root mean square deviation (RMSD) = 13.6199. With superimposition, the RMSD may give different values from 1.2 to 16.9999 when focusing location is shifted. Controversially, with protein folding structure alignment (PFSA), the protein 3D conformational comparison is expressed in an alignment table in Table 4. The alignment table explicitly shows the similarity and dissimilarity in local structures. Also, the value of PFSA-S = 0.4712 gives an unambiguous meaning in global structural similarity within normalization scope between one and zero.

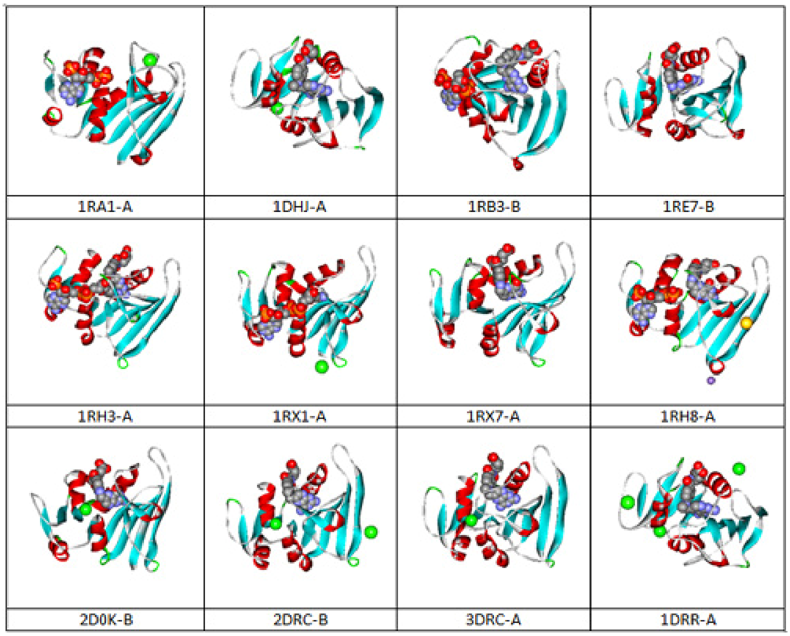


Figure 5. Images of 12 structures of protein enzyme of Dihydrofolate reductase (DHFR).

Table 2. Alignment Table for structural comparison of DHFR (dihydrofolate reductase)

	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000111
	1 2 3 4 5 6 7 8 9 0
Ruler	34567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012
Seq.	SLIAALAVDRVIGMENAMPWNLPADLAWFKRNTLDKPVIMGRHTWESIGRPLPGRKNIILSSQPGTDDRVTWVKSVDEAIAACGDVPEIMVIGGGRVYEQ
1RA1-A	BBBBBBWZQSBVJVAPYHHJBBVAAAAAAAAAJWYHJBBBHAAAAAAJVPYHAAHBBBBBBVPYJVJVAAIBBVAAAAAAAAJWSBAJBBBBBVAAAAAA
1RB3-B	BBBBBBWZQSBWSWYPCYHJBBVAAAAAAAAAJWYHJBBBHAAAAAAJVPYHAAHJBBBBBVPYJVJVAAIBBVAAAAAAAAJWSBAJBBBBBVAAAAAA
2DRC-B	BBBBBBWZQSBVJVAPYHHBBBVAAAAAAAAAAAAHJBBBHAAAAAAJVPYHAAHJBBBBBVPYJVJVAAIBBVAAAAAAAAJWSBAJBBBBBHAAAAAA
3DRC-A	BBBBBBWZQSBWSWYPYHHJBBVAAAAAAAAAAAAHIBBBHAAAAAAJVPYHAAHBBBBBBVPYJVJVAAIBBVAAAAAAAAJWSBAJBBBBBVDAAAAA
2D0K-B	BBBBBBWZQSBVJVAPYHHJBBVAAAAAAAAAAAAHJBBUHAAAAAAJVPYHAAHBBBBBBVJBBHBVAABBBVAAAAAAAAJWSBAJBBBBBVDAAAAA
1DHJ-A	BBBBBBWZQSBWSWYPYHHJBBVAAAAAAAAAAAAHIBBBHAAAAAAJVPYHAAHBBBBBBVPYJVJVAAIBBVAAAAAAAAJWSBAJBBBBBWZAAAAA
1RE7-B	BBBBBBWZQSBVJWYPYHHJBBVAAAAAAAAAAAAHJBBVHAAAAAAJVPYHAAABBBBBBVAJBVJVAAIBBVAAAAAAAAJWSBAJBBBBBWZAAAAD
1DDR-A	BBBBBBWZQSBWSWYPYHAJBBVAAAAAAAADAAAHIBBBHAAAAAAJVPYHAAHBBBBBBVPYBVJVAAJBBVAAAAAAAAJWSBAJBBBBBPZAAAAA
1RX8-A	BBBBBBWZQSBWSW$YHBVABBVAAAAAAAAAAAAHJBBVHAAAAAAJVPYHAAHBBBBBBVPYJVJVAAIBBVAAAAAAAAJWSBAJBBBBBVDAAAAA
1RX7-A	BBBBBBWZQSBBLYAAABVJBBVAAAAAAAAAAAAHIBBVHAAAAAAJVAHHAAHJBBBBBVAHJVJVAAIBBVAAAAAAAAJWSBAJBBBBBVDAAAAA
1RX1-A	BBBBBBWZQSBVJW$YHBVABBVAAAAAAAAAAAAHJBBVHAAAAAAJVAHHAAHJBBBBBVPYJVJVAAIBBVAAAAAAAAJWSBAJBBBBBVAAAAAA
1RH3-A	BBBBBBWZQSBWYP$ZHJVABBVAAAAAAAAAAAAHJBBBHAAAAAAJVPYHAAHBBBBBBVPYJVJVAAABBBVAAAAAAAAJWSBVJBBBBBWZAAAAA

Table 2. (Continued).

	1111111111111111111111111111111111111111111111111111111
	`        1         2         3         4         5`
Ruler	3456789012345678901234567890123456789012345678901234567
Seq.	FLPKAQKLYLTHIDAEVEGDTHFPDYEPDDWESVFSEFHDADAQNSHSYCFEILE
1RA1-A	AAAAJVJBBBBBBBBBBBBVJBHHJBBVAAJBBVJBBBBVJWZAJVAIBBBBBBV
1RB3-B	AAAAJVJBBBBBBBBBBBBVJBHHJBBVAAJBBVJBBBBVJWZAJVAJBBBBBBB
2DRC-B	AAAAJVJBBBBBBBBBVJBVJVHHJBBVAAJBBVJBBBBWSWZAJVAIBBBBBBV
3DRC-A	AAAABVJBBBBBBBBBBBBVJVPYJBBVAAJBBVJBBBBVJWZAJVAIBBBBBBB
2D0K-B	AAAABVJBBBBBBBBBBBBVJBHHJBBVAAJBBVJBBBBVJWZAJVAIBBBBBBV
1DHJ-A	AAAABVJBBBBBBBBBBBBVJVPYJBBVAAJBBVJBBBBHJWZAJVAJBBBBBBB
1RE7-B	AAAAJVJBBBBBBHBBBBBVJBPYJBBVAAJBBVJBBBBVJWZAJVAIBBBBBBV
1DDR-A	AAAABVJBBBBBBBBBBBBHJBPYJBBVAAJBBVJBBBBVJWZAJVAIBBBBBBB
1RX8-A	AAAAJVJBBBBBBVHBBBBAJBHHBVJVAAJBBVJBBBBVJWZAJVAIBBBBBBV
1RX7-A	AAAAJVJBBBBBBBBBBBBVJVPYJBBVAAJBBVJBBBBVJWZAJVAIBBBBBBV
1RX1-A	AAAAJVJBBBBBBBBBBBBAJVPYBVJVAAJBBVJBBBBHJWZAJVAIBBBBBBV
1RH3-A	AAAAJVJBBBBBBVJBBBBVJUPYJBBVAAJBBVJBBBBWSWZAJVABBBBBBBV

The names of conformers are listed on left column. The amino acid sequence and rule for number of residue are listed on top rows. The protein folding shape code (PFSC) for each conformer is listed following the structure name. The typical □-helices are remarked with red color, the □-strands with blue color and the tertian fragments with black. Also, these analogue structures with secondary structure are remarked with pink color.

Table 3. Structural Similarity Score of DHFR (dihydrofolate reductase)

PDB-ID	PFSA-S	# Identity	# Analogy
1RA1-A	1.000000	155	0
1RB3-B	0.980645	146	8
2DRC-B	0.972581	144	9
3DRC-A	0.962903	141	11
2D0K-B	0.956452	143	7
1DHJ-A	0.948387	138	12
1RE7-B	0.945161	139	10
1DDR-A	0.940323	136	13
1RX8-A	0.933871	135	13
1RX7-A	0.927419	134	13
1RX1-A	0.916129	133	12
1RH3-A	0.893548	131	10

The left column lists the PDB ID; PFSA-S: protein folding structure alignment score; # Identity: number of identical shapes; # Analogy: number of similar shapes.

Table 4. Alignment table for structural comparison between 1BGE-B and 2GMF-A. The similarity score PFSA-S = 0.4712

<table>
<tr><td>Seq. 1BGE-B
2GMF_A Seq.</td><td>11
6</td><td>PQSFLLKCLEQMRKVQADGTALQETLCATHQLCHPEELVLLGHALGIPQPPLSSCSSQALQLMGCLRQLHSGLFLYQGLLQALAGISPELAPTLDTLQLD
HAAAAAAAAAAAAAAAAAAAAAAAAAADDABBVJVAAAAAAAAAABBBBBBBVAAHAAJBBVAAAAAAAAAAAAAAAAAAAAJWYPYAAAAAAAAAAAAA
+++++++***|**^*||||||||||||**|||*++*|||||*+++|||++++*+++++**+***+^*|*^***|||||||||*^^^*|||||||||||||
-------VJVADJBVAAAAAAAAAAAAAAABBU--HAAAAAJ---BBB----W-----SV-AHJ-BVAJBVJVAAAAAAAAAABW$ZAAAAAAAAAAAAA
-------PSPSTQPWEHVNAIQEARRLLNLSRD--TAAEMNE---TVE----V-----IS-EMF-DLQEPTCLQTRLELYKQGLRGSLTKLKGPLTMMAS</td></tr>
<tr><td>Seq. 1BGE-B
2GMF_A Seq.</td><td>101
83</td><td>TTDFAINIWQQMEDLGMAPATMPAFTSAFQRRAGGVLVASNLQSFLELAYRALRH
AAAAAAAAAAAAAAAHHHIUIUHBBVAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
|||*||*****|*++^^^+++++||**||||||||||||||*++++++++^**^*
AAADAAHJVJVAJ--BBB-----BBWYAAAAAAAAAAAAAAH--------BVPXH
HYKQHCPPTPETS--CAT-----QIITFESFKENLKDFLLVI--------PFDCW</td></tr>
</table>

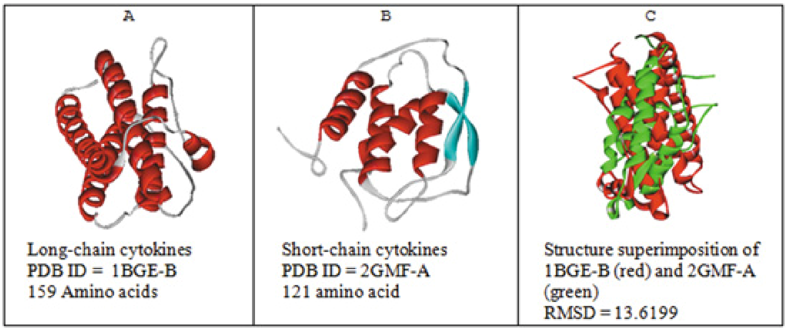


Figure 6. Structural images of protein granulocyte-colony stimulating factor (G-CSF) of 1BGE-B (PDB ID) and protein granulocyte-macrophage colony-stimulating factor (GM-CSF) of 2GMF-A (PDB ID).

Comparison of Proteins with Diverse Degrees in Homology

The comparison of protein structures is a very challenging task, especially for proteins with diverse structural homology. For instance, there is no easy way to superimpose two protein structures together, and specific emphasis of one portion of structures during the superposition may lead to problems for comparing similar structures in the rest of the proteins because they may orient toward different directions in geometric space. In fact, an individual turn point in protein may overshadow the entire similarity comparison between two structures. Moreover, it is more difficult to develop a consistent procedure for comparing the proteins that, on one hand, has the sensitivity for distinguishing different conformers of same protein and, on other hand, is able to take into account of a variety of factors for comparison of different proteins, including the differences in size, sequence homology as well as topological order of secondary structure. Indeed, the structural comparison becomes much more difficult if two proteins belong to different categories in protein classification, such as belonging to different families, superfamilies, folds or classes. However, the PFSA provides a unique process to compare protein structures with diverse homology.

A set of 111 protein structures, for example, with diverse degree in homology is listed in Table 5 as a benchmark, which is composed of 20 conformers of protein 1KZW and various protein domains that were collected to cover other categories from the structure classification of protein (SCOP). In order to diverge from protein of 1KZW (column A in Table 5), the columns of B, C, D, E and F collected different structures by proteins, families, superfamilies, folds and classes according the categories of SCOP to deviate to 1KZW.

All protein domain structures are compared with the first conformer of 1KZW-1, and each pair of comparison generates the PFSA-S to assess the structural similarity. In order to specify the structural similarity between 1KZW-1 and various protein domains, all of values of PFSA-S are display in Figure 7, where the values of PFSA-S are distributed in the same order of columns in Table 5. In Figure 7, the values of PFSA-S for comparisons of 1KZW-1 with other 1KZW conformers are displayed in column A. The PFSA-S values for comparison of 1KZW-1 with other protein domains are displayed in columns B, C, D, E and F, respectively

Table 5. A set of proteins with various homologous and heterologous structures from SCOP classification as benchmark for structural comparison of 1KZW-1

Root	Class: All beta proteins					Other Classes
Class	Fold: Lipocalins				Other Superfamilies	
Fold	Superfamily: Lipocalins					
Superfamily	Family: Fatty acid binding protein-like			Other Families		
Family	Intestinal fatty acid binding protein		Other Proteins			
Protein	Human Homo sapiens					
PDB ID with Chain or Model	1KZW-1	1B56-A	1CBI-A	1AVG-H	1AMM-1	1AUA-A
	1KZW-2	1GGL-A	1EAL-1	1R0U-A	1BWW-A	1AY7-A
	1KZW-3	1HMS-A	1FE3-A	1TXL-A	1D01-A	1DJN-A
	1KZW-4	1JJX-1	1FTP-A	1VPR-A	1F35-A	1E8C-A
	1KZW-5	1KZX-1	1G7N-A	2FR2-A	1JK4-A	1EM8-A
	1KZW-6	1LPJ-A	1GM6-A	2GC9-A	1JS8-A	1G7S-A

Root	Class: All beta proteins					Other Classes
Class	Fold: Lipocalins				Other Superfamilies	
Fold	Superfamily: Lipocalins					
Superfamily	Family: Fatty acid binding protein-like			Other Families		
Family	Intestinal fatty acid binding protein		Other Proteins			
Protein	Human Homo sapiens					
	1KZW-7	1MDC-A	1IFC-A		1M1H-A	1H18-A
	1KZW-8	1O1V-1	1KT7-A		1NKG-A	1HRD-A
	1KZW-9	1O8V-A	1LFO-A		1NLT-A	1J6U-A
	1KZW-10	1PMP-A	1YIV-A		1OK0-A	1KBL-A
	1KZW-11	1TW4-A	2FTB-A		1OLM-A	1O94-A
	1KZW-12	1VYF-A			1QAS-A	1RVV-A
	1KZW-13	2F73-A			1SO9-1	1SW0-A
	1KZW-14	2FS6-A			1TVF-A	1T6T-1
	1KZW-15	2NNQ-A			1YGE-A	1T15-A
	1KZW-16					1V7Z-A
	1KZW-17					1YG6-A
	1KZW-18					2BNH-A
	1KZW-19					2BOD-X
	1KZW-20					2HXM-A
						2JHF-A
						1AMM-A
						1BWW-A
						1D01-A
						1F35-A
						1JK4-A
						1JS8-A
						1M1H-A

Table 5. (Continued).

Root	Class: All beta proteins					Other Classes
Class	Fold: Lipocalins				Other Superfamilies	
Fold	Superfamily: Lipocalins					
Superfamily	Family: Fatty acid binding protein-like			Other Families		
Family	Intestinal fatty acid binding protein		Other Proteins			
Protein	Human Homo sapiens					
						1NKG-A
						1NLT-A
						1OK0-A
						1OLM-A
						1QAS-A
						1SO9-1
						1TVF-A
						1YGE-A
						1DAV-1
						1DLW-A
						1E52-1
						1GRJ-A
						1H9E-1
						1X4T-1
						1Z0J-A
	A	B	C	D	E	F

The relation of protein classification is listed in top 6 rows. The names of protein structures are listed in 6 columns on right, and 6 group names, A, B, C, D, E and F, are listed in bottom row.

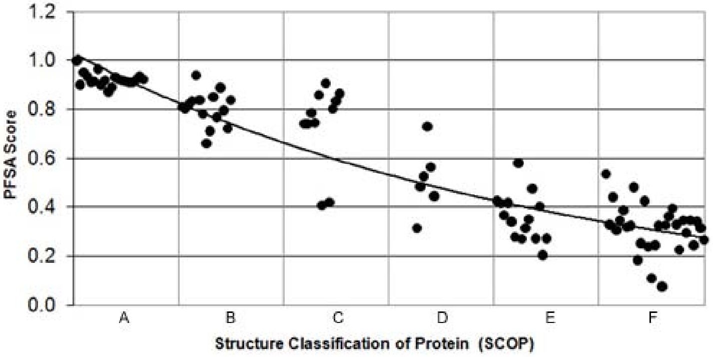


Figure 7. The distribution of PFSA-S for comparison between 1KZW-1 with various protein domain structures belong to different levels of classification. Section A displays the values of PFSA-S for 20 of conformers of protein 1KZW. Referring to 1KZW, section B is for proteins of same family level; section proteins for proteins from different families of same superfamily; section D for proteins from different superfamily of same fold; section E for proteins from different folds of same class and section F for proteins from different classes, which are listed in Table I..

In Figure 7, the PFSA-S value is dispersed between one and zero. The PFSA-S values for comparisons of 1KZW conformers distribute at highest array in column A, and one of these values equals to one that corresponds the comparison of 1KZW-1 with itself. The PFSA-S values for comparisons of 1KZW-1 with different protein structures under same families and superfamilies are spread in the middle array of column B and C. The PFSA-S values for comparisons of 1KZW-1 with protein structures under various classes are distributed in the lower array of column E and F. A polynomial trend line is plotted in Figure 7. Although few of PFSA-S values show false positive or false negative, in overall, the most of values of PFSA-S for comparison of various homologous proteins are coherent with the structure classification of protein (SCOP), i.e. the values of PFSA-S decrease when the degree of structural homology is lower. That demonstrates the values of PFSA-S with overall consistence to protein classification.

Chapter 6

PROTEIN FRAGNMENT SEARCH

With protein folding shape code (PFSC), the digitized description for protein structural conformation is one-dimensional string. It is significant for protein fragment search or probe binding site for drug discovery. First, the search of protein fragment becomes very effective and simple with screening through mass of data from protein database. Second, any specific folding pattern, including typical secondary structure, disrupted secondary structural fragment, β-turn, reverse turn, irregular coil and loop, can be searched because of the completion in description of fording shapes. Third, the targeted fragments can be quantitatively assessed and ranked with structural similarity score, which quickly to lead to the interesting proteins fragments.

Table 6. Fragments with higher structural similarity with 1DUR-A (15-25)

Classification	PDB ID	Score	PFSC	Sequence	Residues
	1DUR-A	1.000	AAAJVABVJBB	KPECPVNCIQE	15-25
Family	1RGV-A	0.818	AAAJVABVJBB	VEECPNEAITP	15-25
Fold	1A9N-A	0.818	AAAJVABVJBB	LASLKSLTYLC	108-118
α/β	2BNH-A	0.818	AAAJVABVJBB	VASQASLRELD	218-228
All β	1SO9-1	0.818	AAAJVABVJBB	PVETQGIKTLT	109-119
Fold	1LOU-A	0.795	AAAJVAJVJBB	LRIRDNVRRVM	79-89
α+β	1QME-A	0.795	AAAJVAJVJBB	LRIRDNVRRVM	79-89
Super	1BQX-A	0.773	AAAJVABBBBB	VEVCPVDCIHE	17-27
Super	1G3O-A	0.773	AAAJVABBBBB	VEECPVDCFYE	17-27
Fold	1UCN-A	0.773	AAAJBHBVJBB	EQKGFRLVGLK	28-38
Fold	2FZC-A	0.773	AAAAAABVJBB	EEVMAEVDILY	216-226
α+β	2C1W-A	0.773	AAAAAABVJBB	NELWDADQNRM	12-22
Super	1F5B-A	0.750	AAABVABBBBB	VEVCPVDCFYE	17-27
Super	1PC4-A	0.750	AAABVABBBBB	VEVCPVDCFYE	17-27
Super	1PC5-A	0.750	AAABVABBBBB	VEVCPVDCFYE	17-27

Classification	PDB ID	Score	PFSC	Sequence	Residues
Super	7FD1-A	0.750	AAABVABBBBB	VEVCPVDCFYE	17-27
Super	7FDR-A	0.750	AAABVABBBBB	VEVCPVDCFYE	17-27
Fold	1L3K-A	0.750	AAAAAAJVJBB	VDKIVIQKYHT	141-151
α+β	1UDX-A	0.750	AAAJVJBBBBB	IARTRVLLYVL	233-243
α+β	2CC6-A	0.750	AAAJVJBWSBB	EDTLDNVVWAE	27-37
α–β	1HRD-A	0.750	AAAJVAJBBBB	LMQQPNMVVAP	358-369
α–β	1V7Z-A	0.750	AAAJVJBBBBB	RVAAGDCVLML	16-26
All α	1Z0J-A	0.750	AAAAAABVIBB	DYADSIHAIFV	136-146

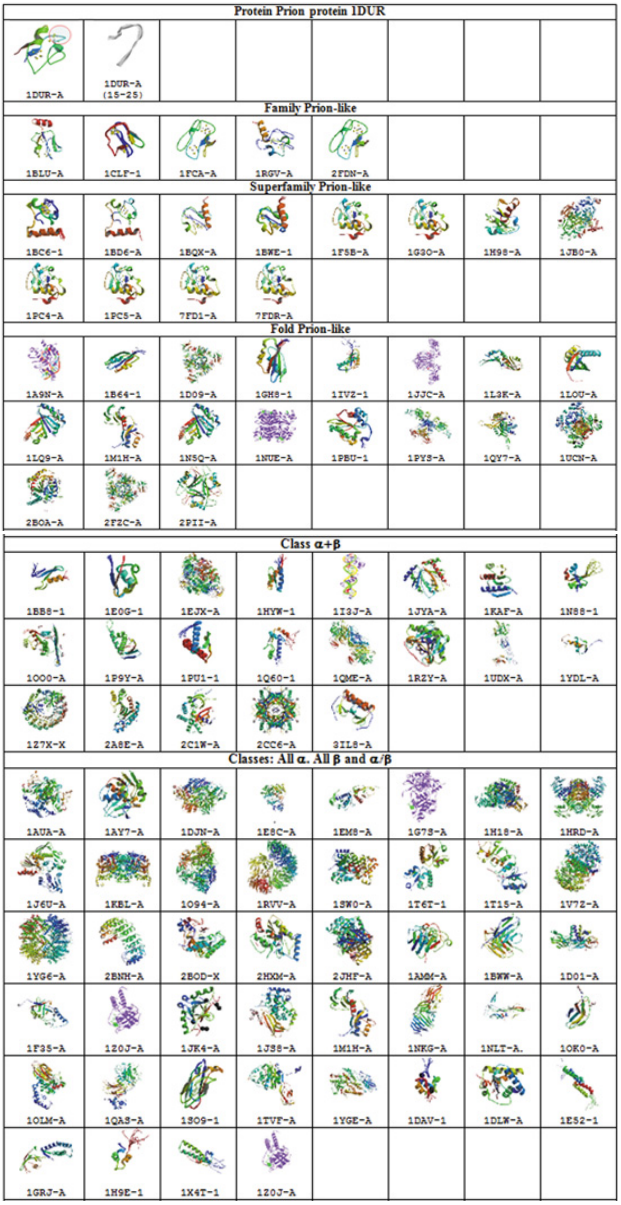


Figure 8. A library of protein 3D structures with various homologies according SCOP classification for searching the fragments with similar folding shape of protein fragment 1DUR-A (15-25).

For example, the fragment IDUR-A (15-25) is queried from a set of protein structures in Figure 8. As a benchmark, this set of protein structures contains 110 proteins with diverse homological character according SCOP, which 5 are belonged to different proteins under same family, 12 different families under same superfamily, 19 different superfamilies under same fold, 21 different folds under same class and 44 different classes. In order to query the fragment of IDUR-A(15-25) of sequence KPECPVNCIQE with folding shape AAAJVABVJBB, all protein structures are converted into one-dimensional PFSC. With screening these structures, a fragment from each protein with similar folding shape is revealed, and some results are listed in Table 6, which lists the targeted PBD ID, location of residues, sequence, PFSC, similarity score and classification. The similarity scores for all targeted structures are presented in Figure 9. It is apparent that four proteins have same folding shape of fragment as IDUR-A (15-25), which have highest score 0.818 and are displayed in Figure 10. Of course, the fragments have same folding shape, but may have unlike sequences. Also, it is noted that the fragments with same folding shape as IDUR-A (15-25) come from different structural classifications, such as one is in same family, one fold, one α/β and another one all β. The results show that he PFSC provides an effective means for data mining protein database. Therefore, the interesting folding fragment is easily discovered.

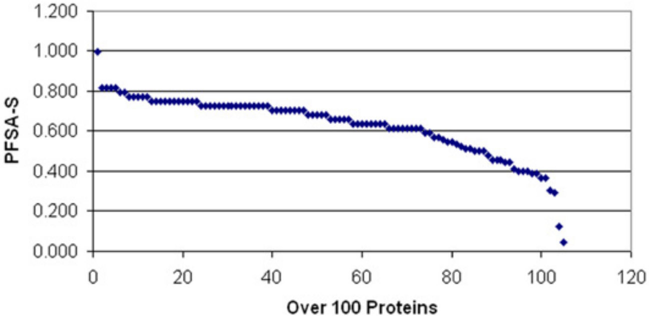
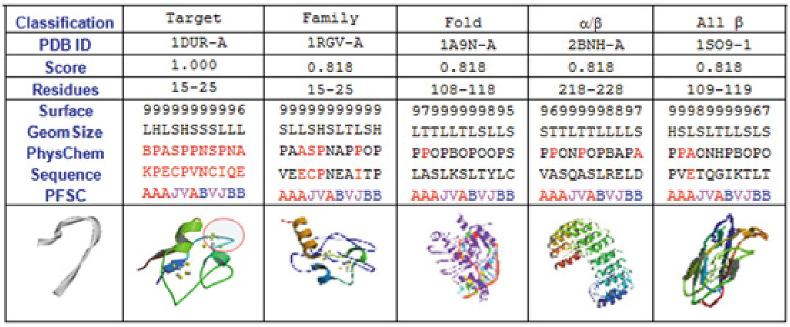


Figure 9. The distribution of PFSA-S for structural similarity of 1DUR-A (15-25) with fragment search through 105 proteins. The set of structures includes proteins n same family, superfamily, fold, classes and other from root.



Classification	Target	Family	Fold	α/β	All β
PDB ID	1DUR-A	1RGV-A	1A9N-A	2BNH-A	1SO9-1
Score	1.000	0.818	0.818	0.818	0.818
Residues	15-25	15-25	108-118	218-228	109-119
Surface	99999999996	99999999999	97999999895	96999998897	99989999967
Geom Size	LHLSHSSSLLL	SLLSHSLTLSH	LTTLLTLSLLS	STTLTTLLLLS	HSLSLTLLSLS
PhysChem	BPASPPNSPNA	PAASPNAPPOP	PPOPBOPOOPS	PPONPOPBAPA	PPAONHPBOPO
Sequence	KPECPVNCIQE	VEECPNEAITP	LASLKSLTYLC	VASQASLRELD	PVETQGIKTLT
PFSC	AAAJVABVJBB	AAAJVABVJBB	AAAJVABVJBB	AAAJVABVJBB	AAAJVABVJBB

Figure 10. Top four of targeted fragments with higher structural similarity with 1DUR-A (15-25).

CONCLUSION

Although more and more protein structures are available in database, and even most of proteins have high resolution in structure, the study of protein complicated structures still is a challenge task for researcher. For single protein, it is not easily to express the folding structural characters clearly from visual modeling with brief statement. For two proteins, no a unique methods is available to compare the similarity with 3D structure superimposition, and even for same method, different procedure may cause different outputs. For multiple proteins, it is hard quantitatively to rank the structures according structural similarity. The successful development of protein folding shape code (PFSC) and protein folding structure alignment (PFSA) approach is a breakthrough for protein structure description. The PFSC provides a complete description of various folding shapes for protein conformation and the 27 PFSC vectors have the integrated relationship and explicit physical meaning. Also, the PFSA is well applied to the protein conformational analysis with appropriate sensitivity while it is able to distinguish the proteins with diverse homology. The PFSC and PFSA are significance in application of protein structure and database. First, it provides a complete deception for protein folding structure conformation. Second, with consistent procedure, it provides the both of alignment table and the similarity score for protein folding structural comparison. Third, with quantitative assessment for protein similarity, the process of protein data mining becomes effectively, especially probe binding site for small molecules, which provides the lead information for possibility of inhibition, activity and toxicity in drug discovery.

REFERENCES

[1] Pauling, L. (1940). A theory of the structure and process of formation of antibodies, *J. Am. Chem. Soc.*, *62*, 2643-2657.

[2] Pauling, L. & Corey, R. B. (1951). Atomic Coordinates and Structure Factors for Two Helical Configurations of Polypeptide Chains. *Proc. Natl. Acad. Sci.* USA*, 37*, 235-240.

[3] Linderstrom-Lang, K. U. (1952). *Protein and Enzyme, Lane Medical Lecture*, No. *VI*, P.58. Stanford Univ. Press, Stanford, California.

[4] Rao, S. T. & Rossman, M. G. (1973). Comparison of super-secondary structures in proteins. *J. Mol. Biol.*, *76*, 241-256.

[5] Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature* (London), *261*, 552-558.

[6] Bernal, J. D. (1958). Molecular Mechanism of Rate Processes in Solids - General Introduction-Structure Arrangements Of Macromolecules, *Discussions of The Faraday Society*., *25*, 7-18.

[7] Gu, J. & Bourne, P. E. (2009). The Worldwide Protein Data Bank in Structural Bioinformatics, second edition, *John Wiley & Sons, Inc., Hoboken*, NJ, 293-303.

[8] Levitt, M. & Warshel, A. (1975). Computer Simulation of *Protein Folding. Nature*, *253*, 694-698.

[9] Bradley, P., Malmstrom, L., Qian B., Schonbrun, J., Chivian, D, Kim, D. E. Meiler, J., Misura, K. M. & Baker, D. (2005). Free modeling Proteins. *Proteins,* 61, 7, 128-134.

[10] Cheng, J. & Baldi, P. A. (2006), Machine Learning Information Retrieval Approach to Protein Fold Recognition, *Bioinformatics*, *22*, 12, 1456-1463 http://folding.stanford.edu/.

[11] Klepeis, J. L. & Floudas, C. A. (2003). ASTRO-FOLD: a combinatorial and global optimization framework for *ab initio* prediction of three-dimensional structures of proteins from the amino acid sequence, *Biophys J.*, *85(4)*, 2119-2146.

[12] Brooks, C. L. III. (2002). Protein and Peptide Folding with Molecular Simulations. *Acc. Chem. Res.*, *35*, 447-454.

[13] Zhang, Y. (2008). Progress and challenges in protein structure prediction, *Curr. Opin. Struct. Biol.*, *18(3)*, 342–348.

[14] Moult, J., et al. (2009). *Critical* assessment of methods of protein structure prediction – Round VIII. *Proteins*, 77 Suppl. *9*, 1-4.

[15] Pauling, L., Corey, R. B. & Branson, H. R. (1951). The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain., *Proc. Natl. Acad. Sci.*, USA, *37*, 205-211.

[16] Eisenberg, D. (2003). The discovery of the α-helix and β-sheet, the principal structural features of proteins. *Proc. Natl. Acad. Sci.*, USA, 100, 1,1207-11210.

[17] Pauling, L. & Corey, R. B. (1951), The Pleated Sheet, A New Layer Configuration of Polypeptide Chains, *Proc. Natl. Acad. Sci.*, USA, *37*, 251-256.

[18] Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963), Stereochmistry of polypeptide chain configurations, *J. Mol. Biol.*, *7*, 95-99.

[19] Venkatachalam, C. M. (1968), Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units, *Biopolymers*, *6*, 1425-1436.

[20] Rose, G. D., Gierasch, L. & Smith, J. A. (1985) Turns in Peptides and Proteins. Advances in Protein Chemistry, New York, *Academic Press*, *37*, 1-109.

[21] Leszczynski, J. F. & Rose, G. D. (1986), Loops in globular proteins: a novel category of secondary structure. *Science*, *234*, 849-855.

[22] Orengo, C. A. & Thornton, J. M. (2005), Protein families and their evolution - A structural perspective, *Annual Review of Biochemistry*, *74*, 867-900.

[23] Park, J. H., Ryu, S. Y., Kim, C. L. & Park, I. K. J. (2001), Protein Classification Comparison Server, *Genome Informatics*, *12*, 350-351.

[24] Hadley, C. & Jones, D. T. (1999), A systematic comparison of protein structure classifications: SCOP, CATH and FSSP, *Structure*, *7(9)*, 1099-1112.

[25] Holm, L. & Sander, C. (1993), Protein structure comparison by alignment of distance matrices, *J. Mol. Biol.*, A(233),123-138.

[26] Gerstein, M. & Levitt, M. (1996), Using iterative dynamic programming to obtain accurate pair-wise and multiple alignments of protein structures in: Proc. Fourth Int. Conf. on Intell. *Sys., for Mol. Biol.,* Menlo Park, CA, AAAI Press., 59-67.

[27] Gibrat, J. F., Madel, T. & Bryant, S. H. (1996), Surprising similarities in structure comparison, *Curr. Opin. Struct. Biol.*, *6*, 377-385.

[28] Singh, A. P. & Brutlag, D, L. (1997). Hierarchical protein structure superposition using both secondary structure and atomic representations, In Proc. Fifth Int. Conf. on Intell. *Sys. for Mol. Biol.* Menlo Park, CA, AAAI Press., 284-293.

[29] Singh, A., Brutlag, D. *3dSearch - Secondary Structure Superposition*, Stanford Bioinformatics Group, http://gene.stanford.edu/3dSearch.

[30] Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng.*, *11*, 9, 739-47.

[31] Krissinel, E. & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, *Acta Crystallogr D Biol Crystallogr.*, *60*, 12, 2256-2268.

[32] Balaji, S., Sujatha, S., Kumar, S. S. C. & Srinivasan, N. (2001). PALI-a database of alignments and phylogeny of homologous protein structures, *Nucleic Acids Res.*, *29*, 61-65.

[33] Choi, I. & Kim, S. H. (2006), Evolution of protein structural classes and protein sequence families, *Proc. Natl. Acad. Sci.*, USA, *103*, 14056-14061.

[34] Hou, J., Sims, G. E., Zhang, C. & Kim, S. H. (2003), A global representation of the protein fold space, *Proc. Natl. Acad. Sci.*, USA, *100*, 5, 2386-2390.

[35] Irving, J. A., Whisstock, J. C. & Lesk, A. M. (2001). Protein structural alignments and functional genomics, *Proteins*, *42*, 378-382.

[36] Sam, V., Tai, C. H., Garnier, J., Gibrat, J. F., Lee, B. & Munson, P. J. (2006), ROC and confusion analysis of structure comparison methods identify the main causes of divergence from manual protein classification, BMC Bioinformatics, BMC *Bioinformatics*, 7, 206.

[37] Fischer, D. & Eisenberg, D. (1996), Protein fold recognition using sequence-derived predictions, *Prot. Sci.*, *5*, 947-955.

[38] Kabsch, W. & Sander, C. (1983), Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, 22, 2577-2637.

[39] Ridchards, F. M. & Kundrot, C. E. (1988). Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure, *Proteins*, *3*, 71-84.

[40] Frishman, D. & Argos, P. (1995). Knowledge-based protein secondary structure, *Proteins, 23*, 566-579.

[41] Sklenar, H., Etchebest, C. & Lavery, R. (1989). Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis, *Proteins*, *6*, 46-60.

[42] Labesse, G., Colloc'h, N., Pothier, J. & Mornon, J. P. (1997). P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins, *Comput. Appl. Biosci.*, *13*, 3, 291-295.

[43] Martin, J., Letellier, G., Marin, A., Taly, J. F., de Brevern, A. G. & Gibrat, J. F. (2005). Protein secondary structure assignment revisited: a detailed analysis of different assignment methods, *BMC Struct. Biol.*, *5*, 17-34.

[44] Fetrow, J. S., Palumbo, M. J. & Berg, G. (1997), Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme, *Proteins*, *27*, 249-271.

[45] Zhang, X., Fetrow, J. S. & Berg, G. (1994). Design of an Auto-associative Neural Network with Hidden Layer Activations that were used to Reclassify Local Protein Structures, *Advances in Protein Chemistry*, In: Crabb, VJ, editor. San Diego, CA: Academic Press, 397-404.

[46] Brevern, A. G., Etchebest, C. & Hazout, S. (2000), Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks, *Proteins*, *41*, 271-287.

[47] Alexandre, G., de Brevern1, Valadié, H., Hazout, S. & Etchebest, C. (2002). Extension of a local backbone description using a structural alphabet: A new approach to the sequence-structure relationship, *Prot. Sci.*, *11*, 2871-2886.

[48] Fourrier, L., Benros, C. & Brevern, A. G. (2004), Use of a structural alphabet for analysis of short loops connecting repetitive structures, *BMC Bioinformatics*, *5*, 58.

[49] Fitzkee, N. C., Fleming, P. J., Gong, H., Panasik N, Street , T. O. & Rose, G. D. (2005). Are proteins made from a limited parts list? *Trends in Biochemical Sciences*, *30*, 73-80.

[50] Govindarajan, S., Recabarren, R. & Goldstein, R. A. (1999). Estimating the total number of protein folds. *Proteins*, *35(4)*, 408-414.

[51] Yang, J. (2008),Comprehensive description of protein structures using protein folding shape code, *Proteins, 713*, 1497-1518.

[52] Yang, J. (2010). *Quantitative and Qualitative Assessment for Protein Folding Structural Similarity*, (Submitted).

[53] Needleman, S. B. & Wunsch, C. D. (1970). A general method applica-ble to search for similarities in the amino acid sequenc-es of two proteins, *J Mol Biol.*, *48*, 442-453.

[54] Crescenzi, O., Tomaselli, S., Guerrini, R., Salvadori, S., D'Ursi, A. M., Andrea, P. & Temussi, Picone, D. (2002). Solution structure of the Alzheimer amyloid β-peptide (1–42) in an apolar microenvironment, *Eur. J. Biochem.*, *269*, 5642–5648.

[55] Tomaselli, S., Esposito, V., Vangone, P., van Nuland, N. I. J., Alexandre, M. J. J., Bonvin, Guerrini, R., Tancredi, T., Temussi, P. A. & Picone A. (2006), The α-to-β Conformational Transition of Alzheimer's Aβ-(1–42) Peptide in Aqueous Media is Reversible: A Step by Step Conformational Analysis Suggests the Location of b Conformation Seeding, *ChemBio. Chem.*, *7*, 257- 267.

INDEX