國立交通大學

生物資訊所

博士論文

以蛋白質結構字元集研究結構與功能之相關性

A Study of Relationships between Protein Structures and Functions Using a Structural Alphabet

研究生：董其樺

指導教授：楊進木 教授

中華民國九十八年九月

以蛋白質結構字元集研究結構與功能之相關性

A Study of Relationships between Protein Structures and Functions Using a Structural Alphabet

研究生：董其樺　　Student：Chi-Hua Tung

指導教授：楊進木　　Advisor：Jinn-Moon Yang

國立交通大學

生物資訊所

博士論文

A Dissertation

Submitted to Institute of Bioinformatics

College of Biological Science and Technology

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

PhD

in

Bioinformatics

September 2009

Hsinchu, Taiwan, Republic of China

中華民國九十八年九月

以蛋白質結構字元集研究結構與功能之相關性

研究生：董其樺　　　　　　　　　　指導教授：楊進木博士

國立交通大學 生物資訊所 博士班

摘　　要

過去幾年，生物功能與系統網路的相關研究發展逐漸加快。由於結構基因體學技術愈漸成熟，蛋白質資料庫所紀錄之結構數量迅速增加，截至 2009 年七月為止，已有超過五萬八千個蛋白質結構被結晶。然而在此同時，結晶結構已被解出，但是尚無法立即明瞭其生物性功能的蛋白質也隨之日漸增加。因此，現今非常急需發展有效率之生物資訊方法，以研究新結晶之蛋白質的結構同源性與演化分類。

針對上述議題，前人提出了數個方法，其中心思想是將蛋白質的局部結構片段，根據 C_α 三度空間座標資訊轉換成一級編碼之結構字元集，藉此研究蛋白質結構相似性與功能分類。為了研究結構與功能之間的關係，我們發展了一系列創新的研究，包括以 kappa-alpha 角度為基礎之結構字元集以及類 BLOSUM 之計分陣列，發現局部結構資訊比胺基酸序列更具有演化上的保留性。

我們將此創新的結構字元集與計分陣列進一步發展為蛋白質快速搜尋比對與功能分類之工具：3D-BLAST 及 fastSCOP。3D-BLAST 以 BLAST 為搜尋引擎，可以快速尋找同源結構蛋白質，藉以分析新結晶結構，並且具有 BLAST 之特性，包括可信賴的統計基礎和快速有效之搜尋能力。我們亦提供 fastSCOP 網頁服務，用以快速辨認結構功能性區域與演化分類。fastSCOP 結合了 3D-BLAST 與結構比對工具，在快速搜尋 SCOP 資料庫後，再確定結構相似度，並調整功能性區域之範圍，最後輸出演化上分類。

我們的研究結果證實，以 kappa-alpha 角度為基礎之結構字元集可代表蛋白質局部片段。而 3D-BLAST 與 fastSCOP 在辨識新結晶結構之演化分類與功能推測的應用上，是為有用且可信之工具服務。3D-BLAST 和 fastSCOP 的網址分別為 http://3d-blast.life.nctu.edu.tw/及 http://fastscop.life.nctu.edu.tw/。

A Study of Relationships between Protein Structures and Functions Using a Structural Alphabet

Student: Chi-Hua Tung　　Advisor: Dr. Jinn-Moon Yang

Institute of Bioinformatics
National Chiao Tung University

Abstract

In the past few decades, the knowledge about biological function and systems has grown rapidly. As structural genomics research provides structural models in genome-wide strategies, the number of protein structures in the Protein Data Bank (PDB) is rapidly rising; as of as of 7-July-2009, there were more than 58,000 proteins. Besides, the accumulating known protein structures with unknown or unassigned functions emphasize the demand of effective bioinformatics methods with which to annotate the structural homology or evolutionary family.

To address the anterior issues, some approaches have been proposed to encode the 3D local structural fragments based on C_a coordinates into a 1D representation based on several letters, also called as 'structural alphabets'. In order to make a study of current structure–function gap, we developed a series of research, including a novel kappa-alpha plot derived structural alphabet and a novel BLOSUM-like substitution matrix, and explored the structure information based on the fact that the local structure is generally more evolutionary conserved than the amino acid sequence.

We have utilized the theory of structural alphabet to rapidly compare protein structure, homologs search (3D-BLAST) and SCOP superfamily assignment (fastSCOP). We present a novel protein structure database search tool, 3D-BLAST, that is useful for analyzing novel structures and can return a ranked list of alignments. This tool has the features of BLAST (for example, robust statistical basis, and effective and reliable search capabilities). In addition, we propose a web server, named fastSCOP, which rapidly identifies the structural domains and determines the evolutionary superfamilies of a query protein structure. fastSCOP server uses 3D-BLAST to scan quickly a large structural classification database and the top ten different superfamilies of protein domains are obtained from the hit lists. And then, a detailed structural alignment tool is adopted to align these top ten structures to refine domain boundaries and to identify evolutionary superfamilies.

With the encouraging results shown, kappa-alpha plot derived structural alphabet is adopted to develop represent the backbone fragments and the 3D-BLAST and fastSCOP is robust and can be a useful server for recognizing the evolutionary classifications and the protein functions of novel structures. 3D-BLAST and fastSCOP are available at http://3d-blast.life.nctu.edu.tw/ and http://fastscop.life.nctu.edu.tw/, respectively.

Acknowledgment

離鄉背井至新竹求學，包含碩博士班的生涯，至今已七個寒暑。我的學生身份即將告一段落，回頭看看這七年冷暖甘苦的時光，有許多貴人為我的研究與生活伸出援手，無私無償地給我幫助，在此我必須誠摯賦予感激之意。

首先感謝楊進木老師在我這幾年的研究所生涯的指導，無論是在我難過低潮時，老師總會悉心指引我方向，提供各種方法來解決我的困境。又或是在我研究順利成果豐碩時，老師總會真心地拍拍我的肩膀給我一聲讚賞與鼓勵。我不會忘記在老師想到什麼很有趣的研究點子時，那種眉飛色舞、眼光閃耀的神情。對於研究未知的問題充滿好奇、熱情，又能保持細心、嚴謹、堅持與執著，感謝老師樹立了如此的榜樣，我將持續學習。

感謝熊昭教授、周三和教授、許聞廉教授、黃鎮剛教授、王雯靜教授與薛佑玲教授，在學位考試期間擔任我的口試委員，不辭辛勞趕來且在考試中不吝賜教，無論是報告或論文的優缺點都一一為我指點，提出很多很重要的建議跟意見。我從口試委員身上看到了所謂頂尖的學術研究人才，所應該具備的條件。這將是我未來走上研究領域後，繼續努力的目標。

實驗室的夥伴們，包括還在努力求學或已畢業的學長、學弟妹，亦是我感激的對象。同屬於結構生物組的成員章維、力仁，感謝你們和我一起討論研究，為了將研究做得更好，我們都會相互批評和指教，這比只自己一個人做研究要好得多。俊辰學長在撰寫研究計畫、文稿修飾或學問知識上，給我許多幫助和教導。彥甫與凱程分別和我一起管理機房和實驗室帳目，有他們的協助使得管理的工作事半功倍。學弟妹們，彥修、怡馨、峻宇、彥超和韋帆，陪我一起談天、娛樂、運動和歡笑，為我的研究生活注入不一樣的活力和緩衝。已畢業的學弟登凱、右儒、一原與振寧，他們在資訊方面所留下的技術，使我在撰寫程式網頁或管理機器上的能力大為增加。其他實驗室夥伴亦提供給我有形無形的幫忙，我仍誠摯感謝，請原諒我在此未能一一答謝。

此外感謝我外宿的房東伯伯和阿姨，在這五年多來，他們幾乎算是我在新竹的父母，給我生活上非常非常多的照顧，無論是住宿環境、飲食或人生建議。他們讓我不需擔心住宿花費，使我省下許多費用。當然也要感謝當初是碩班同學建宏的介紹引薦。

最後最要感謝的是我的家人以及女友。我的爸爸媽媽以及阿姨對我無盡的關愛和付出，弟弟妹妹對我的關心及體貼，每每回家總是能讓我疲憊的心感到滿滿的溫暖，讓我可重新恢復活力。我的女友王璇陪伴我三年多，期間我們分隔兩地聚少離多，見面的次數少到可用紙筆紀錄，但女友依舊不離不棄，給了我許多在空中的甜蜜和鼓勵，也使我的人生更增添了不同的回憶和意義。

將此論文獻給我已過世的爺爺，希望爺爺能對我感到驕傲。

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

In the past few decades, the knowledge about biological function and systems has grown rapidly. There are many approaches to address this large scale of fields, such as genomics (DNA sequences), structural genomics (protein structures) and proteomics (protein expression and interactions). The rapidly increasing rate of new protein structure arising from structural genomics requires the need for methods to rapidly and reliably infer the molecular and cellular functions of these proteins. As structural genomics research provides structural models in genome-wide strategies [1-3], the number of protein structures in the Protein Data Bank (PDB) is rapidly rising [4]; as of June-2009, there were more than 58,000 proteins. Besides, the accumulating known protein structures with unknown/unassigned functions emphasize the demand of effective bioinformatics methods with which to annotate the structural homology or evolutionary family.

Many sequence and structure alignment methods have been developed to discover homologs of newly determined structures [5]. Protein sequence database similarity search programs, such as BLAST and PSI-BLAST [6, 7], are effective computational tools for identifying homologous proteins. However, these approaches are often not reliable for detecting homologous relationships between distantly related sequences. Many other detailed protein structure alignment methods, such as DALI [8], CE [9], MAMMOTH [10], and VAST [11], have also been developed, and these methods compare two known structures, typically based on the Euclidean distance between corresponding residues rather than the distance between amino acid "types" used in sequence alignments. These tools often require several seconds to align two proteins. At this speed, it would take one day to compare a single protein structure with all of those in the PDB. Recently, however, approaches such as ProtDex2 [12] and ProteinDBS [13] have been proposed to search protein structures more quickly by mapping a structure into indexes for measuring the distance of two structures. Other fast search tools, including TOPSCAN [14], SA-Search [15], and YAKUSA [16], describe protein structures as one-dimensional (1D) sequences and then use specific sequence alignment methods to align two structures. Many of these methods have been evaluated based on the performance of two structure alignments but not on the performance of the database search. To our knowledge, none of these methods provides a function analogous to the E-value of BLAST (probably the most widely used database search tool for biologists) with which to examine the statistical significance of an alignment "hit". This current structure-function gap clearly demonstrates the need for more powerful bioinformatics techniques to identify the structural homology or family of a query protein using known protein structures.

To address the anterior questions, many approaches have been proposed to encode the 3D local structural fragments based on Cartesian coordinates into a 1D representation based on several letters, also called as 'structural alphabets' [17-24]. The structural alphabet represents advantageous local structure and has been used to (i) compare/analyze 3D structures [25-27], (ii) predict protein 3D structures from amino acid sequences [17, 19], (iii) reconstruct the protein backbone [21], and (iv) loop modeling [28].

There is other methods use regular secondary structure information in their algorithms. By linear encoding local protein structures, Ramachandran Sequential Transformation (RST) [29] has been proposed and applied to develop efficient protein similarity search tools, SARST [29] and *i*SARST [30]. These tools encode 3D protein structures into two-dimensional Ramachandran maps [31] and transform them into 1D text letters (Ramachandran codes). In addition, RST has been demonstrated suitable to detecting homologs with circular permutations (CPs) in proteins [32].

In order to make a study of current structure–function gap, we developed a series of research and explored the structure information based on the fact that the local structure is generally more evolutionary conserved than the amino acid sequence [33]. Accordingly, we have utilized the theory of structural alphabet to compare protein structure, homologs search [34, 35] and family assignment [36]. Moreover, many sequence-based methods can be applied to mine biologic meanings quickly from protein structures based on this 23-state structural alphabet. However, to the best of our knowledge, structural alphabet has not been used to discover structural motifs in proteins. Therefore, this 23-state structural alphabet can be adopted to develop multiple structure alignment and structure pattern/motif search methods.

One of the important topics in the biological data mining is discovery of frequent patterns in a set of DNA or protein. These patterns usually aim to share biological meanings. Various pattern discovery algorithms use aligned sequences or multiple sequence alignment (MSA) as an input such as PRINTS [37], PROSITE [38], and Pfam [39]. Besides, TEIRESIAS [40], PRATT2 [41] and a specific pattern growth approach [42] are applied to directly identify frequent patterns from unaligned biological sequences without aligning them.

Although pattern discovery approaches with unaligned sequence are more efficiency and less computationally intensive, it may provide the less biological meanings.

However, many of the most functional and evolutionary relationships between homologous protein are so distinct that they cannot be clearly detected through MSA and are evident only by pairwise/multiple structure comparison of the 3D structures. Because of multiple structure alignment is computationally intensive, it makes more efficient in multiple structure alignment based on encoding 3D structure to 1D structural alphabet sequence. Therefore, the application of structural alphabet not only obtains more efficient in multiple structural alignments but also acquires more biological function and meanings in finding structure pattern/motif.

1.2 Thesis overview

First of all, we developed a novel kappa-alpha plot derived structural alphabet and a novel BLOSUM-like substitution matrix, called structural alphabet substitution matrix (SASM) in Chapter 2. This structural alphabet was valuable for reconstructing protein structures from just a small number of structural fragments and for developing a fast structure database search method. Besides, this SASM matrix was designed to offer the preference of aligning structural segments between homologous structures that share low sequence identity. The aligned score from the SASM matrix provides structural similarity estimates and information on evolutionary distance.

In Chapter 3, we described the theory and results of 3D-BLAST based on structural alphabet and SASM. The 3D-BLAST was used to search protein structure database rapidly for all known homologs of a query (new) structure and return a ranked list of alignments. The results showed that our method enhanced BLAST as a search method, using a new structural alphabet substitution matrix to find the longest common substructures with high-scoring structured segment pairs from an SADB database.

In Chapter 4, structural alphabet and SASM was also applied to rapidly identify the structural domains and determine the evolutionary superfamilies of a query protein structure. The web server we built was named as fastSCOP. fastSCOP was the cooperative integration in 3D-BLAST (a fast structural database search tool) and MAMMOTH (a fast detailed structural alignment tool); the former is required for efficiency and the latter for accuracy.

Chapter 5 presented our current studies about Space-Related Pharmamotif (SRP) in interacting site of protein. The SRP is defined as a set of space-related structural motifs that prefers a set of similar protein sub-site structures consistently interact with ligand, DNA or peptide. We demonstrated preliminary results of SRP discovery and motif search. These results mainly illustrated the feasibility of studying SRP. Finally, Chapter 6 described some conclusions and future perspectives.

Chapter 2

Kappa-alpha Plot Derived Structural Alphabet and Structural Alphabet Substitution Matrix

2.1 Introduction

A major challenge facing structural biology research in the post-genomics era is to discover the biologic functions of genes identified by large-scale sequencing efforts. As protein structures increasingly become available and structural genomics research provides structural models in genome-wide strategies [1], proteins with unassigned functions are accumulating, and the number of protein structures in the Protein Data Bank (PDB) is rapidly rising [4]. The current structure-function gap highlights the need for powerful bioinformatics methods with which to elucidate the structural homology or family of a query protein by known protein sequences and structures.

The three-state secondary elements, namely α-helix, β-sheet, and coils, are rather crude for predicting protein structure, and it is not possible to make use of these elements in three-dimensional (3D) reconstruction without additional information. Many approaches have been proposed to replace three-state secondary structure descriptions with various local structural fragments, also known as a 'structural alphabet' [17-23], which can redefine not only regular periodic structures but also their capping areas. Such studies have described local protein structures according to various geometric descriptors (for example, C_α coordinates, C_α distances, α or φ, and ψ dihedral angles) and algorithms (for example, hierarchical clustering, empirical functions, and hidden Markov models [HMMs] [18]). Many of these methods involve protein structure prediction; an exception is the SA-Search tool [15], which is based on C_α coordinates and C_α distances, and which adopts a structural alphabet and a suffix tree approach for rapid protein structure searching.

To address the above issues, we have developed a novel kappa-alpha (κ, α) plot derived structural alphabet and a novel BLOSUM-like substitution matrix, called SASM (structural alphabet substitution matrix), for BLAST [6], which searches in a structural alphabet database (SADB). This structural alphabet is valuable for reconstructing protein structures from just a small number of structural fragments and for developing a fast structure database search method called 3D-BLAST. This tool is as fast as BLAST and provides the statistical significance (*E*-value) of an alignment, indicating the reliability of a hit protein structure. For the purposes of scanning a large protein structure database, 3D-BLAST is fast and accurate and is useful for the initial scan for similar protein structures, which can be refined by detailed structure comparison methods (for example, CE [9] and MAMMOTH [10]).

2.2 (κ, α)-map cluster and structural alphabet

For coding the structural alphabet and calculating the substitution matrix, a pair database of structurally similar protein pairs with low sequence identity was obtained from SCOP 1.65 [43]. Of 2051 families in four major classes (all α, all β, α+β, and α/β) with <40% sequence homology to each other, we excluded a number of problem entries, including poor-quality structures, entries with residue numbering problems, and small-sized families (i.e., number of domains <2). We selected 674 structural pairs (i.e., 1348 proteins) based on the following criteria: (1) one pair was selected for each family, and one extra pair was selected for a family having >15 domains; (2) pairs must have <40% sequence identity; (3) pairs must have rmsd <3.5 Å, with >70% of aligned resides included in the rmsd calculation. In total, these protein pairs had an average sequence identity of 26% (462 pairs below 30% identity), an average rmsd of 2.3 Å, and average aligned residues of 90% (207,492 aligned residues out of 230,915 residues). The amino acid composition of these 1348 proteins was similar to that of proteins in the Swiss-Prot database.

2.2.1 (κ, α)-Map

A structure fragment (five residues long) was defined by the (κ, α)-pair angles as shown in Figure 2.1. The κ angle, ranging from 0° to 180°, of a residue *i* is defined as a bond angle formed by three C_α atoms of residues *i – 2*, *i*, and *i + 2*. The α angle, ranging from –180° to 180°, of a residue *i* is a dihedral angle formed by the four C_α atoms of residues *i – 1*, *i*, *i + 1*, and *i + 2*. A specific series of structural fragments, called the (κ, α) map, represents a protein structure. Therefore, each protein structure may form a specific (κ, α)-map distribution as shown in Figure 2.2.

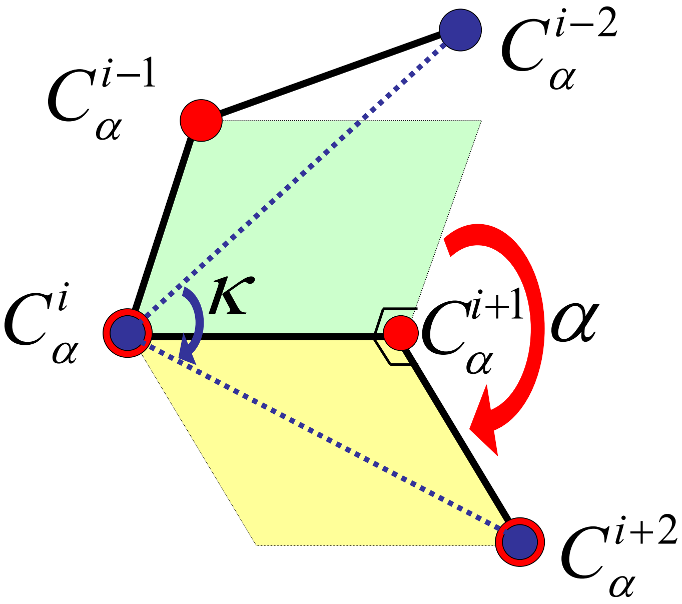


Figure 2.1 Definition of the kappa (κ) and alpha (α) angles.

To code the structural alphabet and calculate the substitution matrix we selected 674 structural pairs (1,348 proteins), which are structurally similar and with low sequence identity, from SCOP based on two criteria: pairs must have rmsd under 3.5 Å, with more than 70% of aligned resides included in the rmsd calculation; and pairs must have under 40% sequence identity. The accumulated (κ, α)-map matrix (Figure 2.3) consists of 225,523 protein fragments derived from 1348 proteins. When the angles of (κ, α) are divided by 10°, this matrix has 648 cells (36*18). The fragment frequency of each cell in this matrix is unbalanced because the protein structures are significantly conserved with regard to α-helix (82,843 segments) and β-strand structures (52,371 segments). Of these helix segments, 71.1% (58,897 segments) are located in four cells that contain 22,310, 15,736, 13,013, and 7,838 segments.

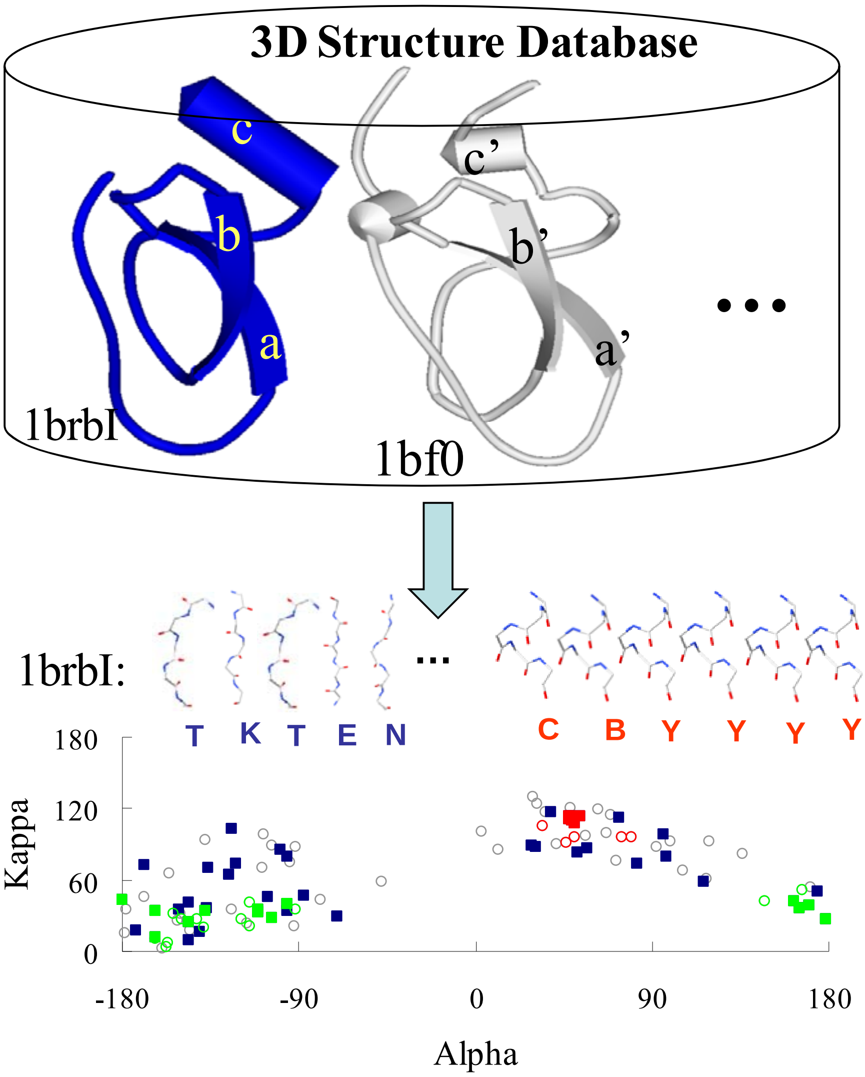
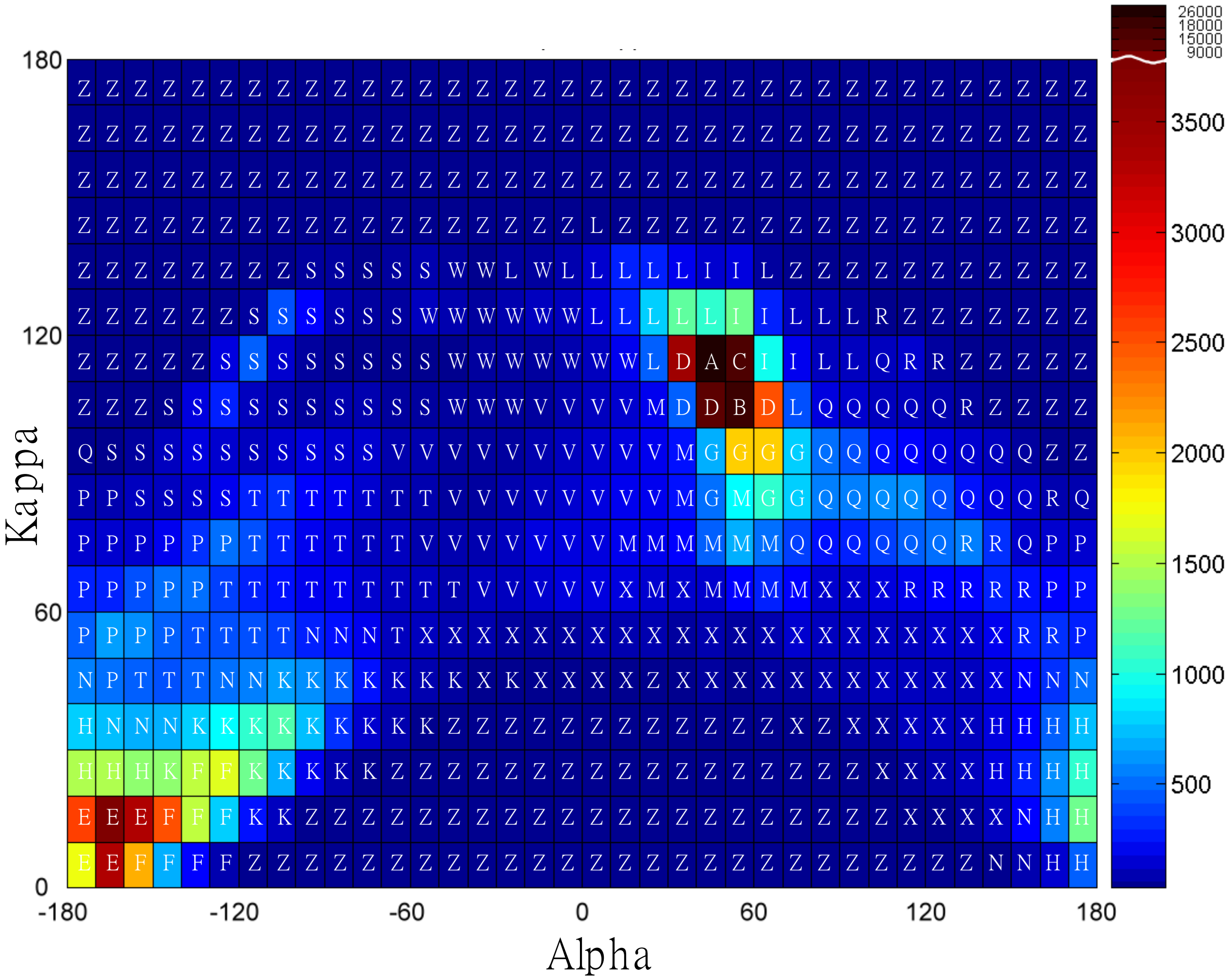


Figure 2.2 The (κ, α) distribution map of 1brbI (square) and 1bf0 (circle).

In the study, the structural distance of a pair of 5-mer protein segments *i* and *j* is determined from the rmsd value of the five C_α atom positions, and is given as follows:

$$\left\{\sum_{k=1}^{5}\left[(X_k-x_k)^2+(Y_k-y_k)^2+(Z_k-z_k)^2\right]/5\right\}^{1/2}$$

Where (X_k, Y_k, Z_k) and (x_k, y_k, z_k) denote the coordinates of the *k*th C_α atom of segments *i* and *j*, respectively. The structural distance is also used to define the intra-segment and inter-segment distances.



Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z
Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z
Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z
Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	L	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z
Z	Z	Z	Z	Z	Z	Z	Z	S	S	S	S	S	W	W	L	W	L	L	L	L	L	I	I	L	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z
Z	Z	Z	Z	Z	Z	S	S	S	S	S	S	W	W	W	W	W	W	L	L	L	L	L	I	I	L	L	L	R	Z	Z	Z	Z	Z	Z	Z
Z	Z	Z	Z	Z	S	S	S	S	S	S	S	S	W	W	W	W	W	W	W	L	D	A	C	I	I	L	L	Q	R	R	Z	Z	Z	Z	Z
Z	Z	Z	S	S	S	S	S	S	S	S	S	S	W	W	W	V	V	V	V	M	D	D	B	D	L	Q	Q	Q	Q	Q	R	Z	Z	Z	Z
Q	S	S	S	S	S	S	S	S	S	S	V	V	V	V	V	V	V	V	V	V	M	G	G	G	G	Q	Q	Q	Q	Q	Q	Q	Q	Z	Z
P	P	S	S	S	S	T	T	T	T	T	T	T	V	V	V	V	V	V	V	V	M	G	M	G	G	Q	Q	Q	Q	Q	Q	Q	Q	R	Q
P	P	P	P	P	P	T	T	T	T	T	T	V	V	V	V	V	V	V	M	M	M	M	M	M	Q	Q	Q	Q	Q	Q	R	R	Q	P	P
P	P	P	P	P	T	T	T	T	T	T	T	T	T	V	V	V	V	V	X	M	X	M	M	M	M	X	X	X	R	R	R	R	R	P	P
P	P	P	P	T	T	T	T	N	N	N	T	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	R	R	P
N	P	T	T	T	N	N	K	K	K	K	K	K	K	X	K	X	X	X	X	Z	X	X	X	X	X	X	X	X	X	X	X	X	N	N	N
H	N	N	N	K	K	K	K	K	K	K	K	K	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	X	Z	X	X	X	X	X	H	H	H	H
H	H	H	K	F	F	K	K	K	K	K	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	X	X	X	X	H	H	H	H
E	E	E	F	F	F	K	K	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	X	X	X	X	N	H	H
E	E	F	F	F	F	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	N	N	H	H

Figure 2.3 The distribution of accumulated (κ, α) plot of 225,523 segments derived from the pair database with 1,348 proteins.

2.2.2 Structural Alphabet

We aimed to use the structural alphabet to represent pattern profiles of the backbone fragments by clustering the accumulated (κ, α)-map matrix (Figure 2.3). A nearest-neighbor clustering (NNC) algorithm was developed to cluster 225,523 fragments in the accumulated (κ, α)-map matrix (Figure 2.3) into 23 groups using the following steps and goals: (1) identifying a representative structural segment for each cell in this matrix; (2) clustering 648 representative segments into 23 groups by grouping similar representative segments and restricting the maximum number of segments in a cluster; (3) in each cluster, identifying a representative segment based on the cell weight which is defined as $w_i = (1/S_i)\Big/\sum_{j=1}^{M}(1/S_j)$, where S_i is the number of segments in cell i and M is the number of cells in this cluster; (4)

assigning the representative segment of a cluster to a structural letter (Figure 2.4); (5) obtaining a composition of 23 structural letters that is similar to the 20 common amino acids. We developed an NNC algorithm instead of using a standard clustering algorithm, such as a hierarchical clustering method or a K-means, which is unable to satisfy the factors (2), (3), and (5).

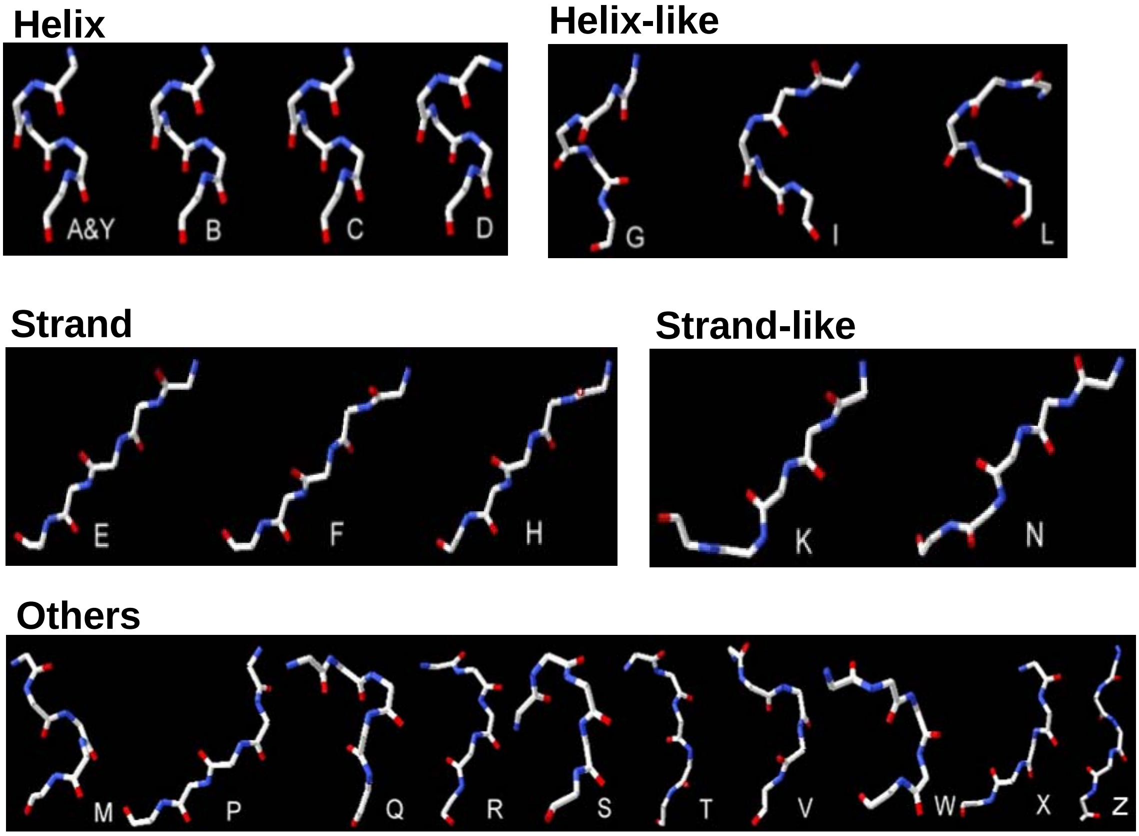


Figure 2.4 The representative 3D fragments of 23 structural alphabets.

3D-BLAST used BLAST as the search method and was designed to maintain the advantages of BLAST. However, 3D-BLAST is slow if the structural alphabet is un-normalized, because the BLAST algorithm searches a statistically significant alignment by two main steps [7]. It first scans the database for words that score more than a threshold value if aligned with words in the query sequence; it then extends each such 'hit' word in both directions to check the alignment score. To reduce the ill effects of using an un-normalized structural alphabet, we set a maximum number (γ) of segments in a cluster in order to have similar compositions for the 23 structural letters and 20 amino acids. The value of γ was set to 16,000 (about 7.0% of total structural segments in the pair database).

According to the restriction parameter *γ*, the cell with the highest number of segments

(22,310) in the accumulated (κ, α)-map matrix should be divided into two subcells by equally separating the κ and α angles: one is located in 100° ≤ κ <105° and 40°≤ α < 45° , and the other is in 105° ≤ κ <110° and 45°≤ α < 50°. These two subcells were labeled as structural letters A and Y, respectively. The NNC method was then applied to cluster the remaining 203,213 fragments into 21 groups. A representative segment of each cell in the accumulated (κ, α)-map matrix was first determined. For each cell, a segment distance matrix (d), stored with the rmsd values by computing all-against-all segments, was created. And the size was $N \times N$, where N is the total number of the segments in a cell. An entry (d_{ij}), which represents the structural distance of segments i and j, is computed by the rmsd of five C_α atom positions and isgiven as

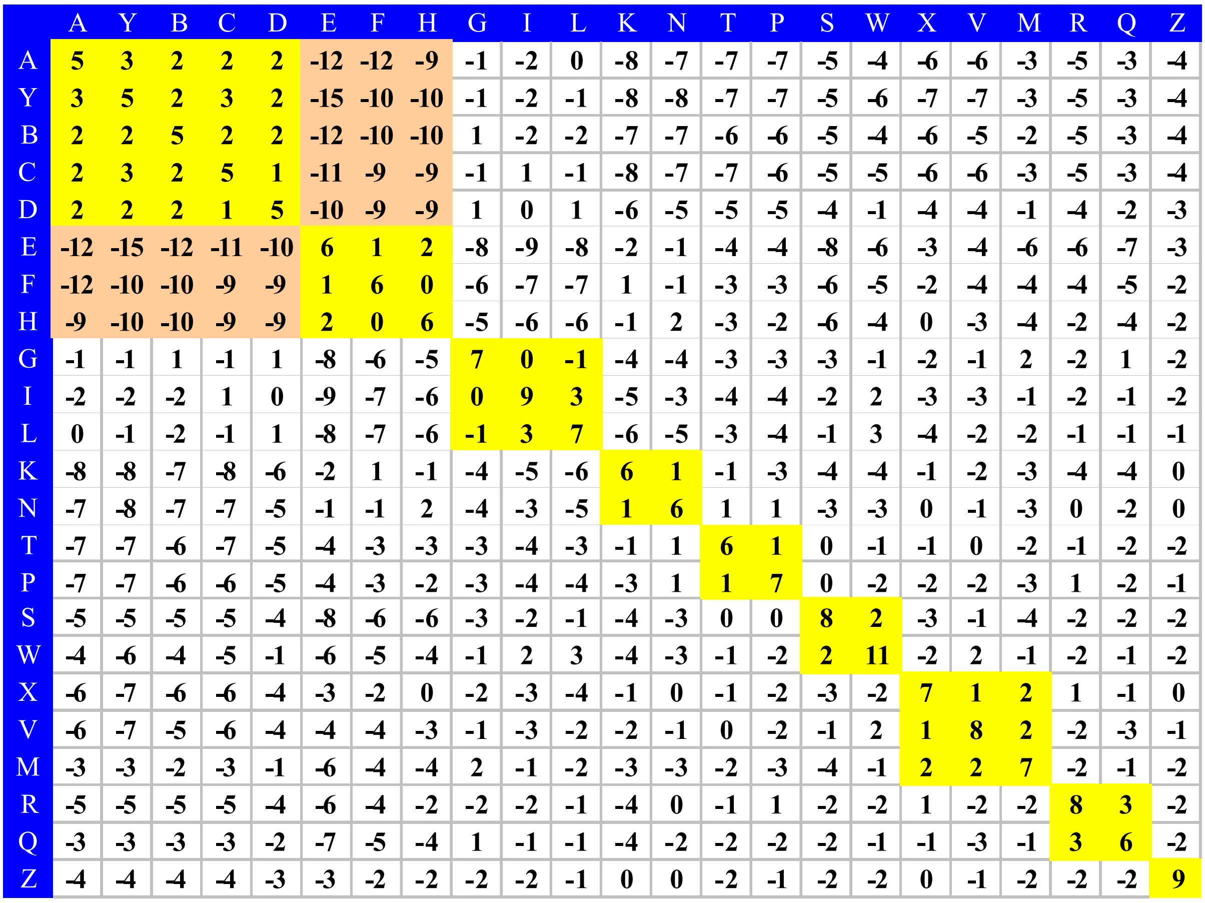
$$\sqrt{\sum_{k=1}^{5}[(X_k - x_k)^2 + (Y_k - y_k)^2 + (Z_k - z_k)^2]/5}$$

where (X_k, Y_k, Z_k) and (x_k, y_k, z_k) are the coordinates of the kth atom of the segments i and j, respectively. For each segment i, the sum of distance (d_i) between the segment i and the other segments in this cell is $\sum_{m=1}^{N} d_{im}$. The segment with the minimum sum of distance is selected as the representative segment of a cell. After the representative segment of each cell is identified, a distance matrix (D) is stored with the rmsd values by computing all-against-all representative segments for these 647 segments. Each entry (D_{ij}, $1 \le i, j \le 647$) is a measure of structural similarity, as defined in Equation 1, between representative segments i and j. In order to ensure that the 3D conformations of the segments clustered in the same group are similar, an rmsd threshold (ε) of the structural similarity is set to 0.5.

Based on the distance matrix D and restriction parameters (ε and γ), the NNC method works as follows: (1) Create a new cluster (C_i, $1 \le i \le 20$) by first selecting an unlabeled cell (a) with the maximum number of segments. Label this cell as C_i. (2) Add an unlabeled cell, which is the nearest neighbor (i.e., a minimum rmsd value in row a of matrix D) of the cell a, into this cluster if this rmsd value is less than ε, and the sum of segments in this cell is less than γ. Label this cell as C_i. Repeat this step until an added cell violates the restriction thresholds, ε or γ. (3) Repeat steps 1 and 2 until the number of clusters equals 21 or all of the cells are labeled. (4) Assign all of the remaining unlabeled cells to a cluster C_{22}. Here, $\varepsilon = 0.95$ Å and $\gamma = 16{,}000$.

Finally, we determined a representative segment and assigned a structural letter for each cluster. For each cell i in a cluster, its sum of distance (D_i) with all of the other cells in the same cluster is equal to $\sum_{m=1}^{N} w_i w_m D_{im}$, where M is the total number of cells in a cluster, w_i is the cell weight, and D_{im} is the structural distance between representative segments i and m of the cells i and m, respectively. The segment with the lowest sum of distance is selected as the representative segment of this cluster. We sequentially assigned a structural letter for each cluster except J, O, and U, since these three letters are not used in BLAST. Figure 2.3 shows the distribution of these 23 clusters and the structural alphabet on 648 cells in the (κ, α) map. Figure 2.4 shows the 3D conformation of each structural segment.

Our new NNC methods, (κ, α) map, and the structural alphabet are easily applied to build new SADB databases from known protein structure databases. We have created several SADB databases derived from PDB, a non-redundant PDB chain set (nrPDB), all domains of SCOP1.69, SCOP1.69 with <40% identity to each other, and SCOP1.69 with <95% identity to each other.



	A	Y	B	C	D	E	F	H	G	I	L	K	N	T	P	S	W	X	V	M	R	Q	Z
A	5	3	2	2	2	-12	-12	-9	-1	-2	0	-8	-7	-7	-7	-5	-4	-6	-6	-3	-5	-3	-4
Y	3	5	2	3	2	-15	-10	-10	-1	-2	-1	-8	-8	-7	-7	-5	-6	-7	-7	-3	-5	-3	-4
B	2	2	5	2	2	-12	-10	-10	1	-2	-2	-7	-7	-6	-6	-5	-4	-6	-5	-2	-5	-3	-4
C	2	3	2	5	1	-11	-9	-9	-1	1	-1	-8	-7	-7	-6	-5	-5	-6	-6	-3	-5	-3	-4
D	2	2	2	1	5	-10	-9	-9	1	0	1	-6	-5	-5	-5	-4	-1	-4	-4	-1	-4	-2	-3
E	-12	-15	-12	-11	-10	6	1	2	-8	-9	-8	-2	-1	-4	-4	-8	-6	-3	-4	-6	-6	-7	-3
F	-12	-10	-10	-9	-9	1	6	0	-6	-7	-7	1	-1	-3	-3	-6	-5	-2	-4	-4	-4	-5	-2
H	-9	-10	-10	-9	-9	2	0	6	-5	-6	-6	-1	2	-3	-2	-6	-4	0	-3	-4	-2	-4	-2
G	-1	-1	1	-1	1	-8	-6	-5	7	0	-1	-4	-4	-3	-3	-3	-1	-2	-1	2	-2	1	-2
I	-2	-2	-2	1	0	-9	-7	-6	0	9	3	-5	-3	-4	-4	-2	2	-3	-3	-1	-2	-1	-2
L	0	-1	-2	-1	1	-8	-7	-6	-1	3	7	-6	-5	-3	-4	-1	3	-4	-2	-2	-1	-1	-1
K	-8	-8	-7	-8	-6	-2	1	-1	-4	-5	-6	6	1	-1	-3	-4	-4	-1	-2	-3	-4	-4	0
N	-7	-8	-7	-7	-5	-1	-1	2	-4	-3	-5	1	6	1	1	-3	-3	0	-1	-3	0	-2	0
T	-7	-7	-6	-7	-5	-4	-3	-3	-3	-4	-3	-1	1	6	1	0	-1	-1	0	-2	-1	-2	-2
P	-7	-7	-6	-6	-5	-4	-3	-2	-3	-4	-4	-3	1	1	7	0	-2	-2	-2	-3	1	-2	-1
S	-5	-5	-5	-5	-4	-8	-6	-6	-3	-2	-1	-4	-3	0	0	8	2	-3	-1	-4	-2	-2	-2
W	-4	-6	-4	-5	-1	-6	-5	-4	-1	2	3	-4	-3	-1	-2	2	11	-2	2	-1	-2	-1	-2
X	-6	-7	-6	-6	-4	-3	-2	0	-2	-3	-4	-1	0	-1	-2	-3	-2	7	1	2	1	-1	0
V	-6	-7	-5	-6	-4	-4	-4	-3	-1	-3	-2	-2	-1	0	-2	-1	2	1	8	2	-2	-3	-1
M	-3	-3	-2	-3	-1	-6	-4	-4	2	-1	-2	-3	-3	-2	-3	-4	-1	2	2	7	-2	-1	-2
R	-5	-5	-5	-5	-4	-6	-4	-2	-2	-2	-1	-4	0	-1	1	-2	-2	1	-2	-2	8	3	-2
Q	-3	-3	-3	-3	-2	-7	-5	-4	1	-1	-1	-4	-2	-2	-2	-2	-1	-1	-3	-1	3	6	-2
Z	-4	-4	-4	-4	-3	-3	-2	-2	-2	-2	-1	0	0	-2	-1	-2	-2	0	-1	-2	-2	-2	9

Figure 2.5 Structural alphabet substitution matrix (SASM).

2.3 Structural Alphabet Substitution Matrix (SASM)

A substitution matrix is the key component of a protein alignment method. In general, a similar underlying mathematical structure is used to construct these matrices [44]. Here, we developed a Structural Alphabet Substitution Matrix (SASM) (Figure 2.5) by applying this mathematical structure to a structural pairing database consisting of 207,492 structural letters derived from 207,492 structural segments based on the aligned residues in the pair database. This SASM matrix was designed to offer the preference of aligning structural segments between homologous structures that share low sequence identity. The aligned score from the SASM matrix provides structural similarity estimates and information on evolutionary distance.

The entry (S_{ij}), which is the substitution score for aligning a structural letter i, j pair ($1 \leq i$, $j \leq 23$), of the SASM matrix is defined as $S_{ij} = \lambda \log_2 \frac{q_{ij}}{e_{ij}}$, where λ is a scale factor for the matrix. q_{ij} and e_{ij} are the observed probability and the expected probability, respectively, of the occurrence of each i, j pair. The observed probability is $f_{ij} \Big/ \sum_{m=1}^{23} \sum_{k=1}^{m} f_{mk}$, where f_{ij} is the total number of letter i, j pairs in these 207,492 structural letters. The expected probability is $p_i p_j$ for $i = j$ and $2p_i p_j$ for $i \neq j$, where p_i is the background probability of occurrence of letter i. The p_i is given as $q_{ii} + \sum_{k \neq i}^{23} q_{ik} / 2$. The substitution score is greater than zero ($S_{ij} > 0$) if the observed probability is greater than the expected probability. By contrast, $S_{ij} < 0$ if $q_{ij} < e_{ij}$. The optimal λ value is yielded by testing various values ranging from 0.1 to 5.0; is set to 1.89 for the best performance and efficiency. The final score S_{ij} is rounded to the nearest integer value.

2.4 Evaluation of (κ, α)-Map and Structural Alphabet

The goal of creating a structural alphabet is to define the 3D structure of fragments of the protein backbone and then represent a protein structure in 3D by a series of structural letters. A structural letter represents pattern profiles of the fragment backbones (five residues long) derived from the pair database; therefore, a protein structure of L residues is described by a structural alphabet sequence of L-4 letters. Here, we used the pair angles, κ (from 0° to 180°) and α (from –180° to 180°) as shown in Figure 2.1, to divide a 3D protein structure into a series of 3D protein fragments.

Figure 2.3 shows the accumulated (κ, α) map matrix (648 cells) of 225,523 3D segments derived from 1348 proteins in the pair database when the κ and α angles are divided by 10°.

The number of 3D segments in each cell ranges from 0 to 22,310, and the color bar on the right side shows the distribution scale. According to the definitions in DSSP, the numbers of α-helix and β-strand segments are 82,482 (36.57%) and 52,371 (23.33%), respectively. In this (κ, α) map, most of the α-helix segments are located on four cells in which the α angle ranges from 40° to 60° and the κ angle ranges from 100° to 120°. In contrast, the κ angle of most of the β-strand segments ranges from 0° to 30°, and the α angle ranges from –180° to –120° or from 160° to 180°. The number of cells having no segments is 183. We observed that most of the 3D segments in a cell have similar conformations; that is, the root-mean-square deviation (rmsd) is less than 0.3 Å on five contiguous C_α-atom coordinates. Moreover, the conformations of 3D segments located in adjacent cells are often more similar than ones in distant cells. These results indicate that the (κ, α) map matrix is useful for clustering these 3D segments and for determining a representative segment for each cluster.

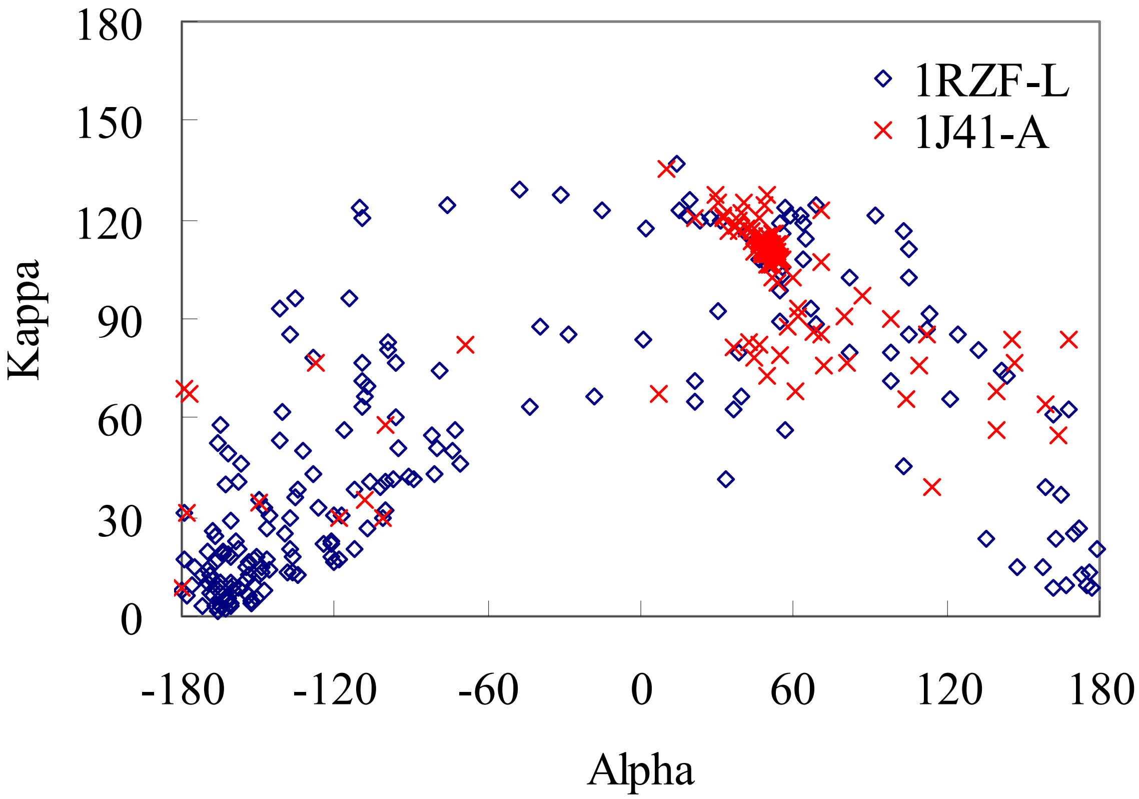


Figure 2.6 The (κ, α) plots of an all-α protein (Protein Data Bank [PDB] code 1J41-A; red) and an all-β protein (PDB code 1RZF-L; blue).

Each structure has a specific (κ, α) plot (Figure 2.6) when governed by these two angles. For instance, a typical (κ, α) plot (blue diamond) of an all-β protein (human anti-HIV-1 GP120-reactive antibody E51, PDB code 1RZF-L [45]) is significantly different from that (red cross) of an all-α protein (human hemoglobin, PDB code 1J41-A [46]). Conversely, two similar protein structures have similar (κ, α) plots.

The (κ, α) plot is similar to a Ramachandran plot, based on the following observations. First, the α-helices are located in very restricted areas, in which α ranges from 40° to 60°, and κ ranges from 100° to 120°. Additionally, β-sheet segments are restricted to some regions in the (κ, α) plot. All residues are fairly restricted in their possibilities in both plots. Second, angles φ and ψ in the Ramachandran plot, denoting a protein structure with a series of 3D positions of amino acids, are widely adopted to develop various structural segments (blocks). Here, the (κ, α) plot was utilized to develop a structural alphabet, which represents a protein structure as a series of 3D protein fragments, each of which are five residues long. The angles φ and ψ represent the position relationship of two contiguous amino acids, whereas the angles κ and α represent the position relationship of five amino acids. These observations indicate that the (κ, α) plot is an effective means of both developing short sequence structure motifs and assessing the quality of a protein structure.

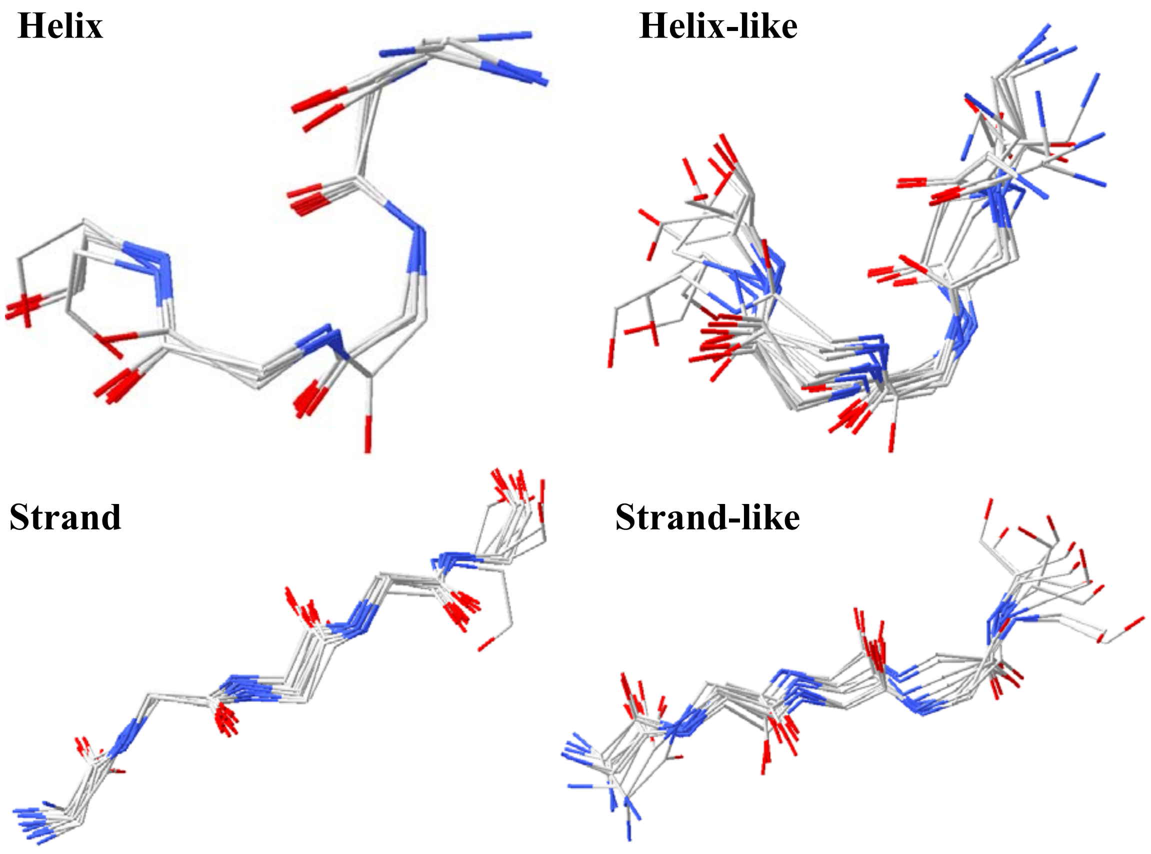


Figure 2.7 The three-dimensional (3D) segment conformations of the five main classes of the 23-state structural alphabet.

A set of representative segments with 23 states and its respective structural letters are identified (Figure 2.7) after performing the NNC method. Here, this 23-state structural alphabet was adopted for both protein structure reconstructions and protein structure database searches. The intra-segment structural distances (blue) are much greater than the inter-segment structural distances (Figure 2.8), and the average rmsd values of these 3D representative segments located in the same (or similar) cluster are frequently below 0.8 Å. The composition of the 23-state structural alphabet resembles that of the 20 amino acids obtained from the pair database. The distribution of the 23-state structural segments is consistent with that of the eight-state secondary structures defined by the DSSP program.

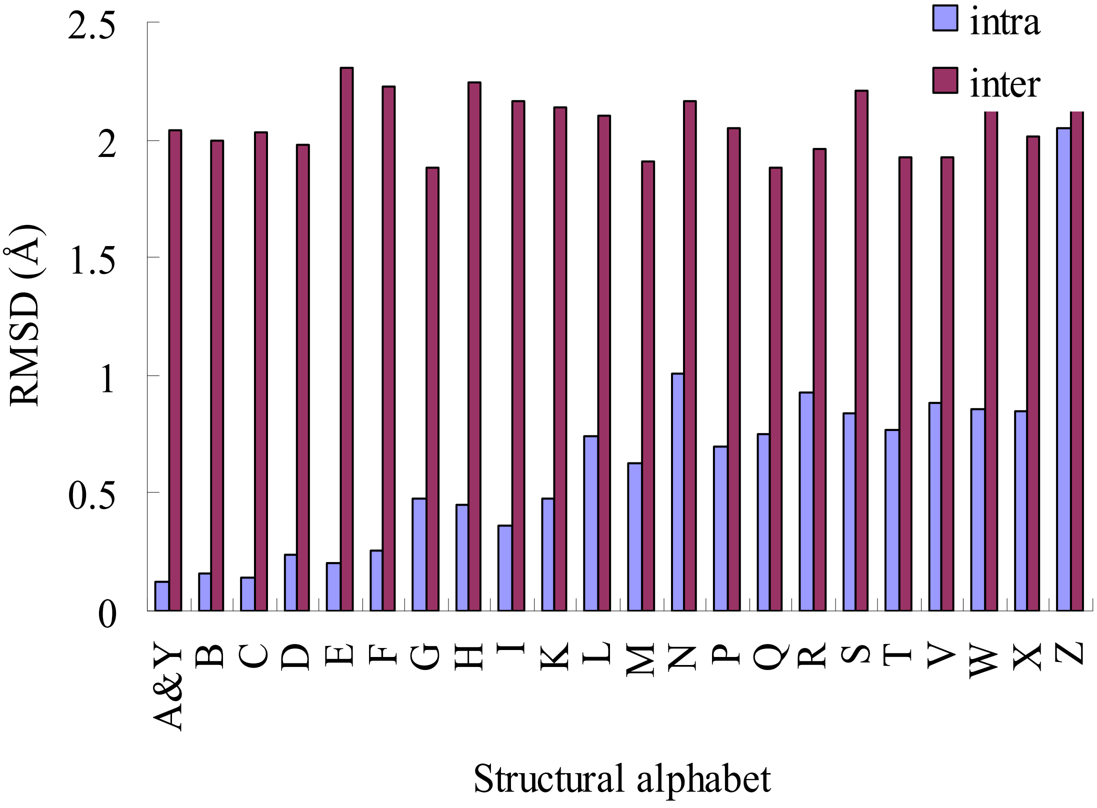


Figure 2.8 The average intra-segment and inter-segment root mean square deviation values of the 23-state structural alphabet.

Based on the (κ, α) plot and a new nearest neighbor clustering, a new 23-state structural alphabet was derived to represent the profiles of most 3D fragments, and was roughly categorized into five groups (Figure 2.7): helix letters (A, Y, B, C, and D), helix-like letters (G, I, and L), strand letters (E, F, and H), strand-like letters (K and N), and others. The 3D shapes of representative segments in the same category are similar; conversely, the shapes of different categories are significantly different. For instance, the shapes of representative 3D segments in the helix letters are similar to each other, as are those in strand alphabets. In contrast, the shapes of helix letters and strand letters obviously differ. The average structural distance (determined from the rmsd value of five continuous C_α atom positions between a pair of 5-mer segments) of inter-segments in both helix and strand letters is less than 0.4 Å (Figure 2.8), and is much less that those of other letters in the structural alphabet. Additionally, most α-helix secondary structures based on the definition of the DSSP program are encoded as helix or helix-like alphabets, and none are encoded as strand or strand-like alphabets (Figure 2.9). Conversely, most β-strand segments are encoded as strand or strand-like letters.

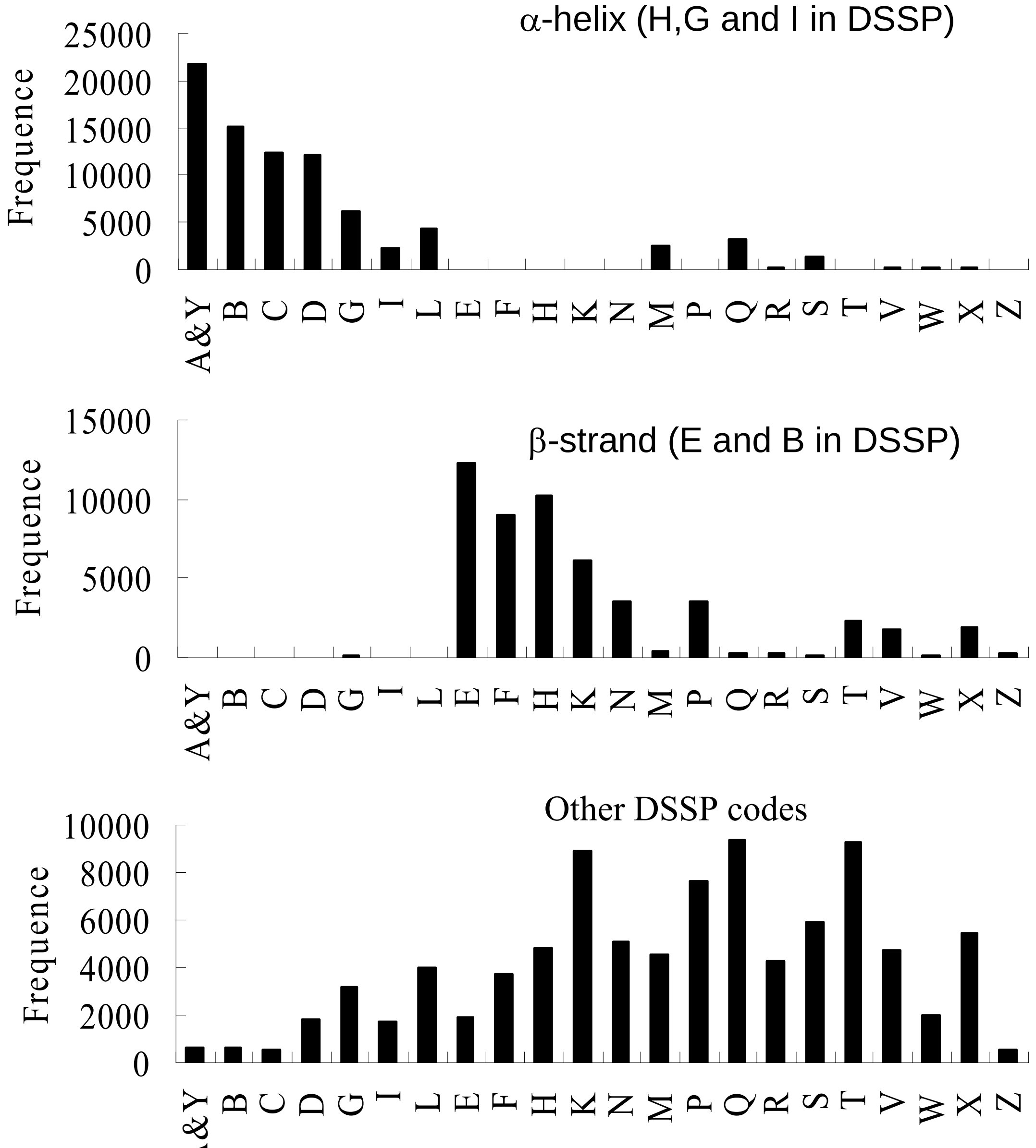


Figure 2.9 The distributions of the 23-state structural alphabet on α-helix, β-strand, and the coil segments defined by the DSSP program.

All residues were fairly restricted in their possibilities in the (κ, α) plot (Figure 2.3). The proportion of cells with 0 segments, which were encoded as structural letter 'Z', was 28.2% (183 cells among 648). Additionally, the numbers of cells and segments with structural letter 'Z' were 272 (42.0%) and 989 (0.4%), respectively. Restated, only 0.44% segments were widely distributed in 41.98% of cells. If the segments of a new protein structure are located on these 41.98% cells, then they may be regarded as poor structural segments. Conversely, five helix letters (A, Y, B, C, and D) and three strand letters (E, F, and H) were located in 7 and 30 cells (Figure 2.3), respectively. The total number of segments located in these 37 (4.4%) cells was 75,477 (33.5%).

The distribution of a structural alphabet is a key determinant of speed in 3D-BLAST. Since the structure database contained high percentages of α-helix and β-strand structures, we restricted the maximum number of structural segments in a cluster for the NNC algorithm to increase the speed of 3D-BLAST. A structural letter, which represents all of the α-helix segments, will occupy 36.57% of total segments without the restriction based on the NNC algorithm. Here, the restriction maximum number of segments was set to 16,000, which is ~7% of the total segments according to the distribution of 20 amino acids. In the structural alphabet, there are 8 letters (the helix and helix-like) for the α-helix structure and 5 letters (strand and strand-like) for the β-strand structure (Figure 2.4). 3D-BLAST is ~64 times faster if the restriction is applied to the NNC method.

In addition, a greedy algorithm and the same evaluation criteria (global-fit score) proposed by Kolodny et al. [21] were used to evaluate the structural alphabet on reconstructing 10 test proteins. This greedy algorithm reconstructed the protein for increasingly larger segments of the protein by using the best structural fragment, i.e. the one whose concatenation yields a structure of minimal rmsd from the corresponding segment in the protein. The experimental results showed that the global rmsd values were from 2.4 Å to 4.5 Å for these 10 proteins and were lightly worse than Kolodny et al. [21] work. In the future, we will enhance the structural alphabet for protein structure prediction.

2.5 Evaluation of SASM

Substitution matrices are the key component of protein alignment methods. We developed a new SASM (Figure 2.5) using a method similar to that used to construct

BLOSUM62 (22) based on a pair database consisting of 674 pairs of proteins. BLOSUM62 is the most commonly used substitution matrix for protein sequence alignment in BLAST. To calculate the preference of structural letters, we prepared this pair database by selecting structurally similar protein pairs having low sequence identity.

The SASM matrix (23*23) offers insights about substitution preferences of 3D segments between homologous structures having low sequence identity. The highest substitution score in this matrix is for the alignment of a letter “W” with a letter “W”, in which the shape of the representative segment is similar to that of β-turns (Figure 2.4), which allows the peptide backbone to fold back and therefore has great significance in protein structure and function [47]. This substitution score is 11 (Figure 2.5). Based on the tool PROMOTIF [48], most of the segments in “W” are β-turns. When two identical structural letters (e.g., diagonal entries) are aligned, the substitution scores are also high. For example, the alignment scores are 9 and 8 when “I” and “S” are aligned with “I” and “S”, respectively. Most of the substitution scores are positive if two structural letters in the same category (e.g., helix letters A, Y, B, C, and D shown in Figure 2.4) are aligned. On the other hand, the lowest substitution score (–15) in this SASM is for the alignment of the “Y” (a helix letter) with the “E” (a strand letter). All of the substitution scores are low when the helix letters (A, Y, B, C, and D) are aligned with strand letters (E, F, and H). The above relationships are in good agreement with biological functions of the relevant structures, showing that the matrix SASM embodies conventional knowledge about secondary structure conservation in proteins.

We compared the SASM matrix and BLOSUM62 [44]. The highest substitution score is 11 for both matrices. In contrast, the lowest score for SASM (–15) is much lower than that for BLOSUM62 (–4). The main reasons for this large difference are that α-helices and β-strands constitute very different protein secondary structures, and the structural letters pertaining to these two types of structure are more conserved than amino acid sequences. Because the gap penalty is an important factor, various combinations of gap penalties were systematically tested for 3D-BLAST and the SASM matrix based on the pair database (1,348 proteins). Here, the optimal values for the open gap penalty and the extended one are 8 and 2, respectively. These results demonstrate that the structural alphabet, SADB and SASM, may be able to more accurately predict protein structures than simple amino acid sequence analyses.

2.6 Reconstructing protein using Structural Alphabet

A greedy algorithm and the evaluation criteria (global-fit score) presented by Kolodny and coworkers [21] were applied to measure the performance of 23-state structural alphabet (structural segments) in reconstructing the α-β-barrel protein (PDB code 1TIM-A [49]) and 38 structures selected from the SCOP95-1.69 set, which comprises 516 proteins. This greedy algorithm reconstructs the protein in increasingly large segments using the best structural fragment, namely the one whose concatenation produces a structure with the minimum rmsd from the corresponding segment in the protein from 23 structural segments. No energy minimization procedure was utilized to optimize the reconstructing structures in this study. The global rmsd values were from 0.58 Å to 2.45 Å, and the average rmsd value was 1.15 Å for these 38 proteins. Figures 2.10A and B illustrate the reconstructed structures of the α-β-barrel protein and ribonucleotide reductase (PDB code 1SYY-A [50]), respectively. The C_α carbon rmsd values were 0.80 Å (1TIM-A) and 0.63 Å (1SYY-A) between the X-ray structures (red) and reconstructed proteins (green). The reconstructed structures are frequently close to the X-ray structures on both α-helix and β-sheet segments, and the loop segments account for the main differences. If all representative segments (465 segments) of the non-zero cells in the (κ, α) plot were considered when reconstructing structures, then the global rmsd values would be in the range 0.35 to 2.32 Å, and the average rmsd value would be 0.94 Å.

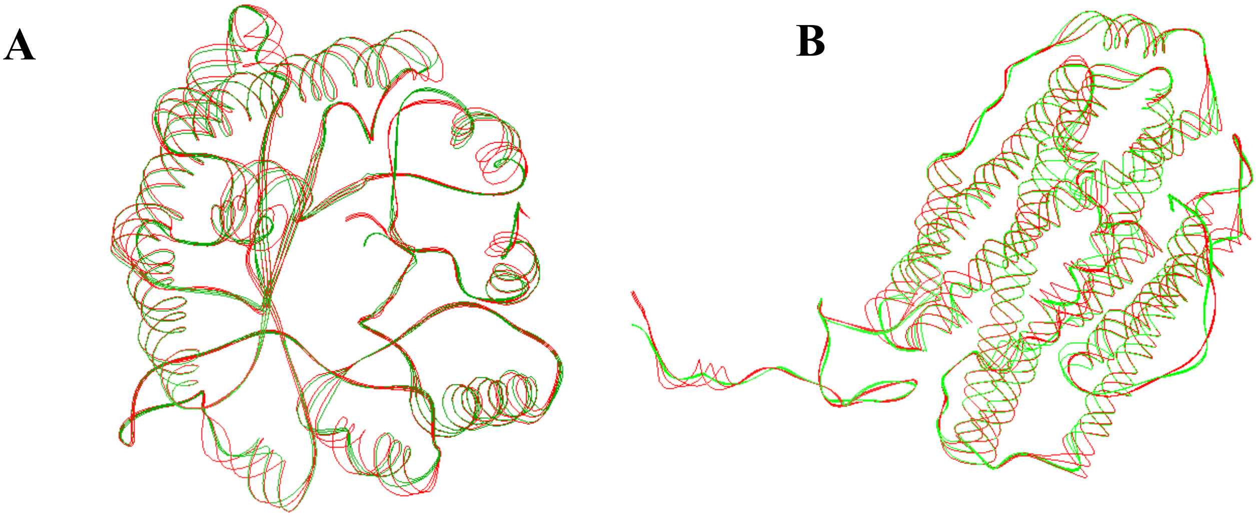


Figure 2.10 Reconstruction protein structures using the 23-state structural alphabet. Reconstruction of the (A) α-β-barrel protein (PDB code 1TIM-A) and (B) ribonucleotide reductase (PDB code 1SYY-A).

The 23-state structural alphabet should be able to represent more biologic meaning than standard three-state secondary structural alphabets. First, the classic regular zones of three-state secondary structures are flexible structures. For instance, α-helices may be curved [51] and more than one-quarter of them are irregular [52], and the φ and ψ dihedral angles of β-sheets are widely dispersed. The proposed 23-state alphabet describes α-helices with eight segments (five helix letters and three helix-like letters) and β-sheets with five segments (Figure 2.7). Figure 2.10 reveals that the 23 structural segments performed well in reconstructing protein structures, particularly in the structure segments of classic α-helices and β-sheets. Second, the three-state secondary structure cannot represent the large conformational variability of coils. Nonetheless, some similar structures can be identified for many of the protein fragments, such as β-turns [47], π-turns, and β-bulges [53]. Here, 10 structural segments in the 23-state alphabet were utilized to describe the loop conformations. An analysis using the PROMOTIF [48] tool reveals that most of the segments (>80%) in the letter 'W' are β-turns.

2.7 Summary

This study demonstrates the robustness and feasibility of the (κ, α) plot derived structural alphabet for developing a small set of sequence-structure fragments and a fast one-against-all structure database search tool. The (κ, α) plot is an effective means of assessing the quality of protein 3D structure.

Future investigations can adopt the (κ, α) plot derived 3D fragment library to develop a small 3D fragment library and predict protein structures. Moreover, many sequence-based methods can be applied to mine biologic meanings quickly from protein structures based on this 23-state structural alphabet.

Chapter 3

Protein Structure Database Search and Evolutionary Classification

3.1 Introduction

Numerous sequence alignment methods (for instance BLAST [6], SSEARCH [54], SAM [55], and PSI-BLAST [7]) and structure alignment methods (for instance, DALI [8], CE [9], and MAMMOTH [10]) have been demonstrated to identify homologs of newly determined structures. Sequence alignment methods are rapid but frequently unreliable in detecting the remote homologous relationships that can be suggested by structural alignment tools; also, although the latter may be useful, they are slow at scanning homologous structures in large structure databases such as PDB [4]. Various tools including ProtDex2 [12], YAKUSA [16], TOPSCAN [14], and SA-Search [15] have recently been developed to search protein structures quickly. TOPSCAN, SA-Search, and YAKUSA describe protein structures as one-dimensional sequences and then use specific sequence alignment methods to replace BLAST for aligning two structures, because BLAST needs a specific substitution matrix for a new alphabet. Many of these methods have been evaluated based on the performance of two structure alignments but not on the performance of the database search. Additionally, none of these methods provides a function analogous to the *E*-value of BLAST (which is probably the most adopted database search tool by biologists) for investigating the statistical significance of an alignment 'hit'.

To the best of our knowledge, 3D-BLAST is the first tool that permits rapid protein structure database searching (and provides an *E*-value) by using BLAST, which searches a SADB database with a SAMS matrix. The SADB database and the SASM matrix improve the ability of BLAST to search for structural homology of a query sequence to a known protein structure or a family of proteins. This tool searches for the structural alphabet high-scoring segment pairs (SAHSPs) that exist between a query structure and each structure in the database. Experimental results reveal that the search accuracy of 3D-BLAST is significantly better than that of PSI-BLAST at 25% sequence identity or less.

3.2 3D-BLAST: Protein structure database search

We designed 3D-BLAST to search a protein structure database for all known homologs of a query (new) structure and for determining its evolutionary classification. Users input a PDB code with a protein chain (for example, 1GR3-A) or a domain structure with a SCOP identifier (for example, d1gr3a_). When the query has a new protein structure, the 3D-BLAST tool enables users to input the structure file in the PDB format. The tool returns a list of protein structures that are similar to the query, ordered by *E*-values, within several seconds. When we searched databases such as SCOP or CATH [56], which are based on structural classification schemes, the evolutionary classification (family/superfamily) of the query protein was based on the first structure in the 3D-BLAST hit list. The output allows users to directly view the superposition of the structures online or download them in the PDB format. The main advantages of 3D-BLAST using BLAST as a search tool include robust statistical basis, effective and reliable database search capabilities, and established reputation in biology.

Figure 3.1 provides an outline of 3D-BLAST. The program quickly scans a structural alphabet sequence database (SADB), which is derived from known protein structures. Here, we used two proteins, 1brb with I chain (blue) and 1bf0 (gray), to describe these steps and concepts. First, we divided a 3D protein structure into 3D fragments, each five residues long, using κ and α angles (Figure 3.1B) as defined in the DSSP program (21). Second, as governed by these angles, each structure in the protein structure database has a specific (κ, α) map distribution (Figure 3.1C), which was then encoded into a corresponding 1D structural alphabet sequence and stored in the SADB database (Figure 3.1D). Third, we used a generalized theory of a substitution matrix to develop a new matrix, SASM, based on 674 structural protein pairs. We then enhanced the sequence alignment tool BLAST, which searches SADB using this SASM, to quickly discover homology structures or evolutionary classifications. The resulting structural alphabet alignment (Figure 3.1E) is reported along with an *E*-value similar to the one assigned by BLAST, and the structure alignment (Figure 3.1F) is also reported. For example, the (κ, α) map distributions (Figure 3.1C) of 1brbI (filled squares) and 1bf0 (open circles) are similar, as are their protein structures (Figure 3.1F). In Figures 3.1C, D, and E, the β-strand structures (green) and helix structure (red) of these two proteins were aligned by 3D-BLAST. The structures are similar even though the amino acid sequence identity is only 21.3%.

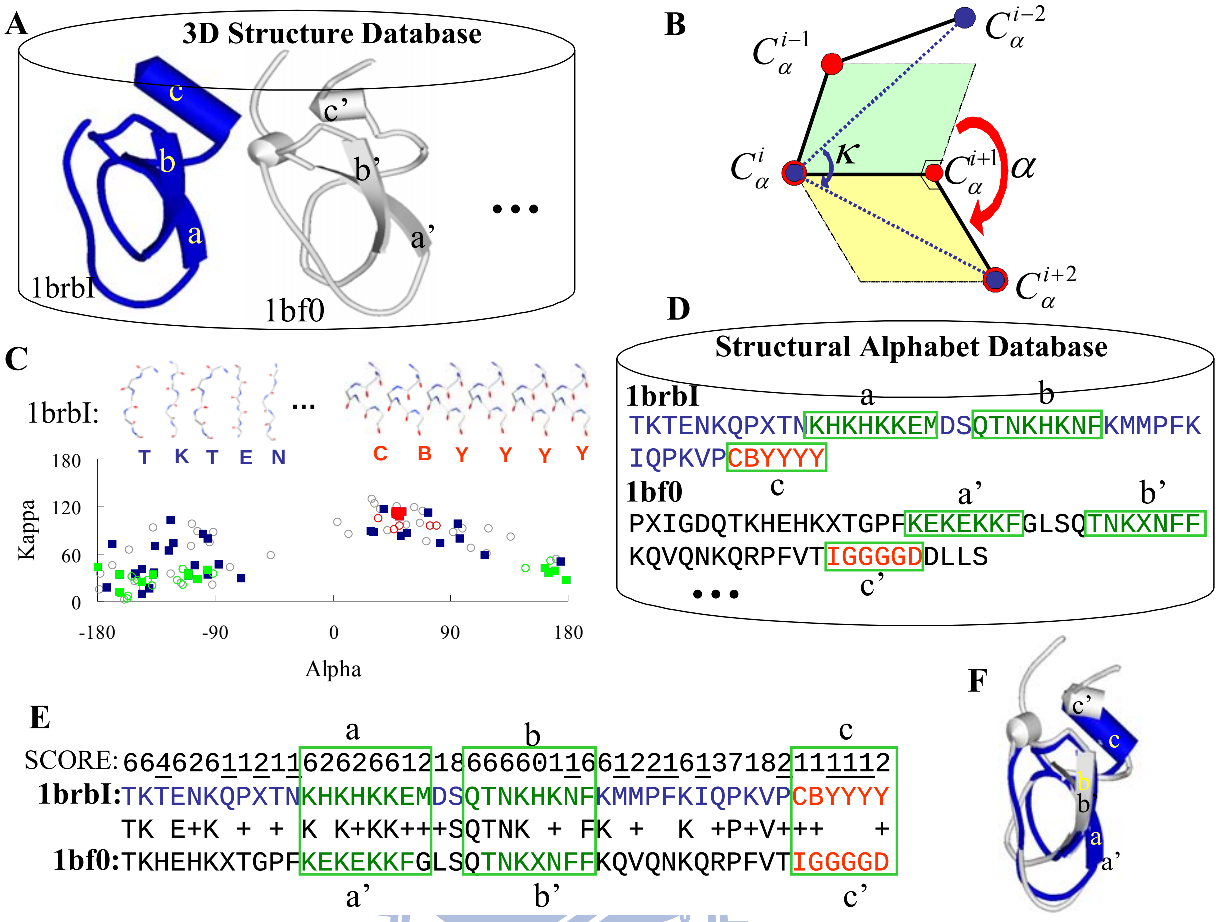

A
3D Structure Database
c
b
a
c’
b’
a’
1brbI
1bf0
• • •
B
C_α^{i-2}
C_α^{i-1}
C_α^{i}
κ
C_α^{i+1}
α
C_α^{i+2}
C
1brbI:
T K T E N
…
C B Y Y Y Y
Kappa
180
120
60
0
-180
-90
0
90
180
Alpha
D
Structural Alphabet Database
1brbI
a
b
TKTENKQPXTNKHKHKKEMDSQTNKHKNFKMMPFK
IQPKVPCBYYYY
1bf0
c
a’
b’
PXIGDQTKHEHKXTGPFKEKEKKFGLSQTNKXNFF
KQVQNKQRPFVTIGGGGDDLLS
• • •
c’
E
a
b
c
SCORE: 66462611211626266121866660116612216137182111112
1brbI:TKTENKQPXTNKHKHKKEMDSQTNKHKNFKMMPFKIQPKVPCBYYYY
TK E+K + + K K+KK+++SQTNK + FK + K +P+V+++ +
1bf0:TKHEHKXTGPFKEKEKKFGLSQTNKXNFFKQVQNKQRPFVTIGGGGD
a’
b’
c’
F
c’
c
b
b
a
a’

Figure 3.1 Stepwise illustration of 3D-BLAST using the protein 1brb chain I as the query protein.

3.3 Datasets and Evaluation Criteria

To evaluate the utility of 3D-BLAST for discovery of homologous proteins and evolutionary classification of a query structure, we selected one query protein set, termed SCOP-894, from SCOP 1.67 and SCOP 1.69, in which the sequence identity is <95%. For evolutionary classification, we considered the first position of the hit list of a query as the evolutionary family/superfamily of this query protein. SCOP-894 contains 894 query proteins from two subsets, SCOP95-1.69 and SCOP95-1.67. The first subset (SCOP95-1.69) contains 516 query proteins that are in SCOP 1.69 but not in SCOP 1.67, and the search database is SCOP 1.67 (11,001 structures). The second subset (SCOP95-1.67) contains 378 query proteins that are in SCOP 1.67 but not in SCOP 1.65, and the search database is SCOP 1.65 (9354 structures). The total number of alignments in SCOP95-1.67 and SCOP95-1.69 is 3,535,812 (378*9354) and 5,676,516 (516*11,001), respectively. Here, a query of 3D-BLAST is a protein sequence with a chain identifier but not a domain sequence.

For comparison with related work on rapid database searching, 3D-BLAST was also tested on a dataset of 108 query domains, termed SCOP-108, proposed by Aung and Tan [12]. These queries, which have fewer than 40% sequence homology to each other, were chosen from medium-sized families in SCOP. The search database (34,055 structures) represents most domains in SCOP 1.65. Finally, the utility of 3D-BLAST for 319 structural genomics targets named as SG-319 was analyzed; the search database was SCOP 1.69, with under 95% identity to each other.

The quality of the 3D-BLAST database search is based on some common metrics, including precision, recall, false positive rate, and receiver operating characteristic (ROC) curve. The precision is defined as A_h/T_h, the recall and false positive rate can be given as A_h/A and $(T_h - A_h)/(T - A)$, respectively, where A_h is the number of true hit structures in the hit list, T_h is the total number of structures in the hit list, A is total number of true hits in the databases, and T is total number of structures in the databases. The ROC curve plots the sensitivity (i.e., recall) against the "1.0 – specificity" (i.e., false positive rate). The average precision is defined as $(A_i\ /\ T_i)/\ A$, where T^i_h is the number of compounds in a hit list containing i correct structures.

3.4 Statistics of 3D-BLAST

A database search method should allow users to examine the statistical significance of an alignment, thereby indicating the reliability of the prediction. 3D-BLAST maintains the advantages of the BLAST tool to provide hit proteins ordered by E-value for fast structural database scanning. 3D-BLAST searches SAHSP, which is similar to the high-scoring segment pair (HSP) in BLAST for protein sequence alignment. Therefore, the statistics of HSPs for analyzing the BLAST algorithm allow us to estimate the E-value of the SAHSP in 3D-BLAST by using the matrix SASM. In BLAST, the statistical significance of a local alignment is accessed with an E-value, which is calculated using the formula $E = Kmne^{-\lambda S}$, where m and n are the lengths of the query and database, respectively, S is the nominal score of the alignment of finding an HSP, and λ and K are statistical parameters based on the scoring system. The E-value is the expected number of chance alignments with a score of S or better. Protein structures and the structural letters are more conserved than protein sequences; thus, as one would expect, the E-values of 3D-BLAST are larger than those of BLAST when the reliable indicators are similar. Here, the λ was set to 1.89 and K was the default value used in BLAST (by testing various values).

To evaluate the accuracy of the *E*-values reported by 3D-BLAST, we submitted shuffled SA sequences as queries and found the number of match sequences with *E*-values below various thresholds. For simplicity, we used the query set SCOP95-1.69 and the respective shuffled queries (516 SA sequences) that represent protein structures, and the search database was SCOP 1.67. Shuffled queries mimic completely random SA sequences, which preserve only the composition basis of a protein structure, using the typical SA composition. The numbers of matches of 516 shuffled queries with *E*-values below e^{-20}, e^{-15}, and e^{-10} are 0, 3, and 326, respectively. On the other hand, the numbers of matches of 516 queries in the SCOP95-1.69 dataset with *E*-values below e^{-20}, e^{-15} and e^{-10} are 8,268, 18,700, and 64,440, respectively. Protein structures and the structural letters are more conserved than protein sequences; thus, as one would expect, the *E*-values of 3D-BLAST are larger than those of BLAST when the reliable indicators are similar.

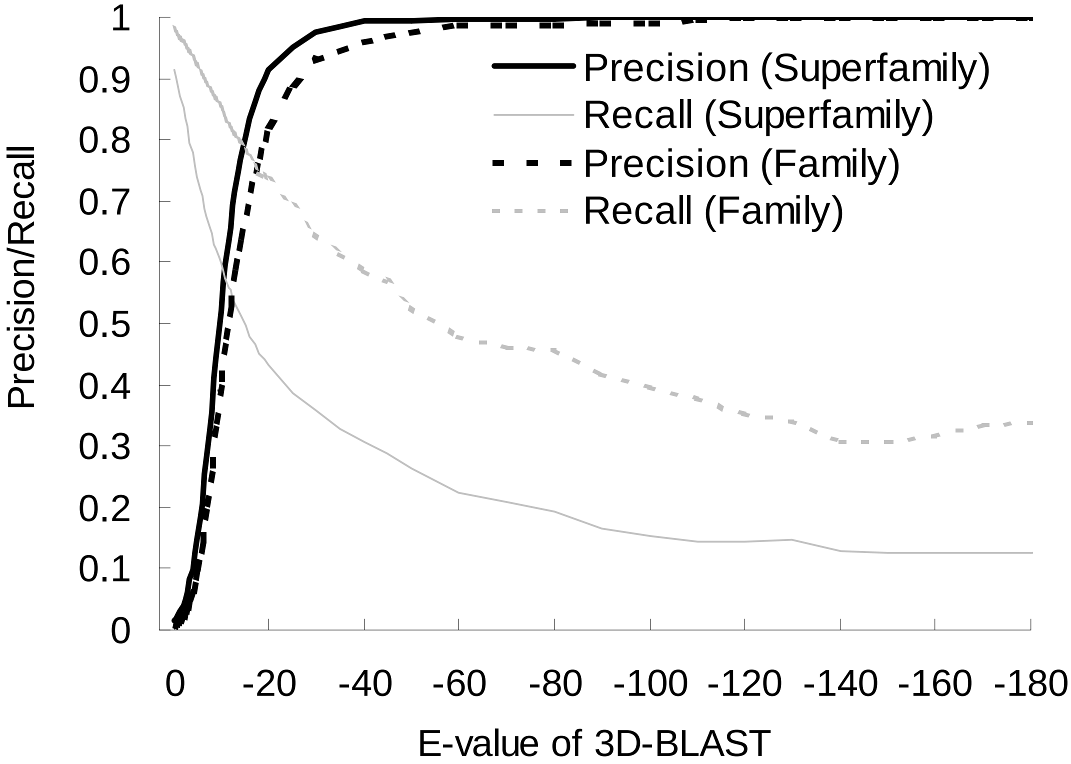


Figure 3.2 3D-BLAST performance with *E*-values: The relationship between precision and recall for structure database search.

Figures 3.2, 3.3, 3.4 and Table 3.1 show the relationships between 3D-BLAST performance and the various *E*-values for SCOP-894. In searching a structural database containing thousands of sequences, generally only a limited number, if any, will be homologous to the query protein structure. Our 3D-BLAST provides cutoff scores to identify highly significant similarity with the query because the biological significance of the high-scoring structures can be inferred on the basis of the similarity score. When a lower *E*-value is used, the proportion of true positives increases for the database search (Figure 3.2) and the rate of correct classification increases for evolutionary classification assignment (Figure 3.3). For structural database searches, the precision is 0.81 and recall is 0.5 if the *E*-value is $<e^{-15}$ (Table 3.1); by comparison, if the cutoff of *E*-value is $<e^{-20}$, the precision is 0.91 and recall is 0.43. For classification assignment, we calculated the relation between the *E*-value of the first hit and the number of correct (thick line) and false (thin line) classification assignments for SCOP-894 (Figure 3.3). If the *E*-value is $<e^{-15}$, 98.53% of 894 protein structures are assigned correct classifications and the coverage is 91.61% (Table 3.1). When the *E*-value is restricted to $<e^{-20}$, 99.60% of the predicted cases are correct and the coverage is 84.23%. When the sequence identity is <25% (229 proteins among 894 proteins), the rate of correct assignments is 92.77% and the coverage is 72.49% if the *E*-value is restricted to $<e^{-15}$.

Table 3.1 3D-BLAST performance with different thresholds of the E-value on structural database searches and automatic SCOP superfamily assignment on the protein query set SCOP-894

	Structural database search			Superfamily assignment [a]			
				894 proteins		Sequence identity < 25% [b]	
Threshold of *E*-value	Recall	Precision	False positive rate	Correct assignment (%)	Coverage [c] (%)	Correct assignment (%)	Coverage (%)
e^{-10}	0.60	0.52	0.0091	96.68%	97.76%	86.32%	92.58%
e^{-15}	0.50	0.81	0.0020	98.53%	91.61%	92.77%	72.49%
e^{-20}	0.43	0.91	0.00056	99.60%	84.23%	97.60%	54.59%
e^{-25}	0.39	0.95	0.00016	99.86%	77.96%	98.94%	41.05%

SCOP-894 consists of 894 query proteins from two subsets, SCOP95-1.67 and SCOP95-1.69. SCOP95-1.67 has 378 query proteins, which are in SCOP 1.67 but not in SCOP 1.65, and the search database is SCOP 1.65. SCOP95-1.69 consists of 516 query proteins, which are in SCOP1.69 but not in SCOP1.67, and the search database is SCOP1.69.

[a] The first rank in the hit list of a query protein is assigned as the superfamily.

[b] The predicted accuracy was calculated from 229 query proteins having <25% sequence identity.

[c] The coverage is defined as *P/T* where *P* is the number of the assigned structures and *T* is total number of structures. For example, *P* is 819 and *T* is 894 if the threshold of *E*-value is set to e^{-15} for the query set SCOP-894.

The proposed 3D-BLAST provides a threshold *E*-value to identify a highly significant similarity with the query. The SASM matrix reveals that the biologic significance of the high-scoring structures can be inferred from the similarity score and the proportion of true positives rises when a lower *E*-value is utilized. Figure 3.4 shows that 3D-BLAST *E*-values correlate with both the Z-scores of CE (blue) and rmsd values (red) of aligned residues. For the 894 query proteins, the Z-scores of CE are >5 and the rmsd values are often <3 Å if the *E*-value is restricted to <e^{-20}. Clearly, if the *E*-values are lowered, the number of true positives and Z-scores of CE increase. These results demonstrate that the *E*-value of 3D-BLAST allows users to examine the reliability of the structure database search and evolutionary superfamily assignments.

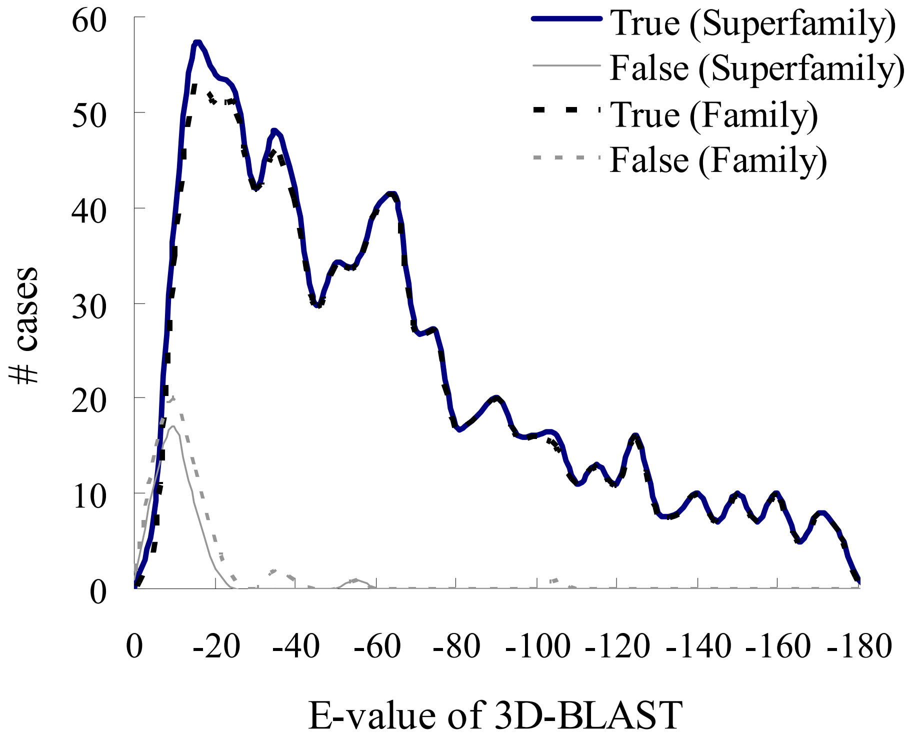


Figure 3.3 3D-BLAST performance with *E*-values: The number of correct and false family/superfamily assignments.

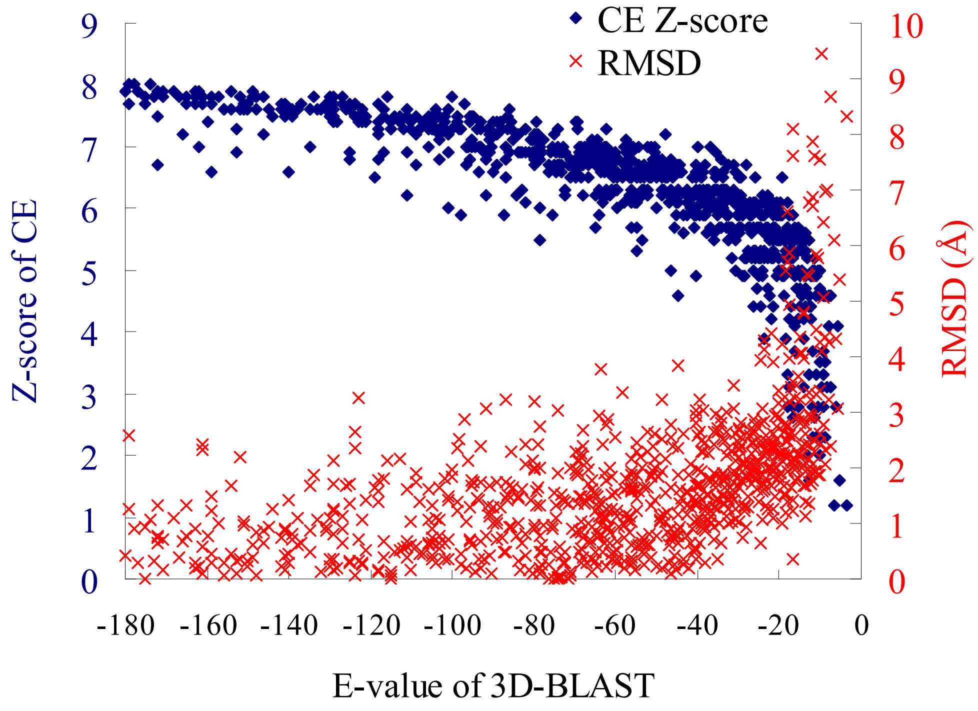


Figure 3.4 3D-BLAST performance with *E*-values: The relationship between 3D-BLAST *E*-values and both Z-Scores of CE and rmsd of aligned residues

Figure 3.5 shows details that *E*-values on the protein query set SCOP95-1.69 correlate strongly with the rmsd values of aligned residues between the query protein and the hit proteins. A total of 22,415 proteins were randomly chosen from the hit lists of 516 query proteins in the SCOP95-1.69 dataset. Among these 22,415 proteins, 27.72% (6,215 structures) had rmsd values below 3.0 Å. If the *E*-value was restricted to under e^{-20}, then 83.52% of hit proteins (2,130 proteins from among 2,549 proteins) had rmsd values less than 3.0 Å, and the average rmsd was 2.37 Å. When the *E*-value was restricted to under e^{-15} and under e^{-10}, then 72.65% (3,984 proteins among 5,487 proteins) and 51.70% (5,742 proteins among 11,106 proteins) of proteins had rmsd values less than 3.0 Å, respectively, and the average rmsd values were 2.85 Å and 3.57 Å.

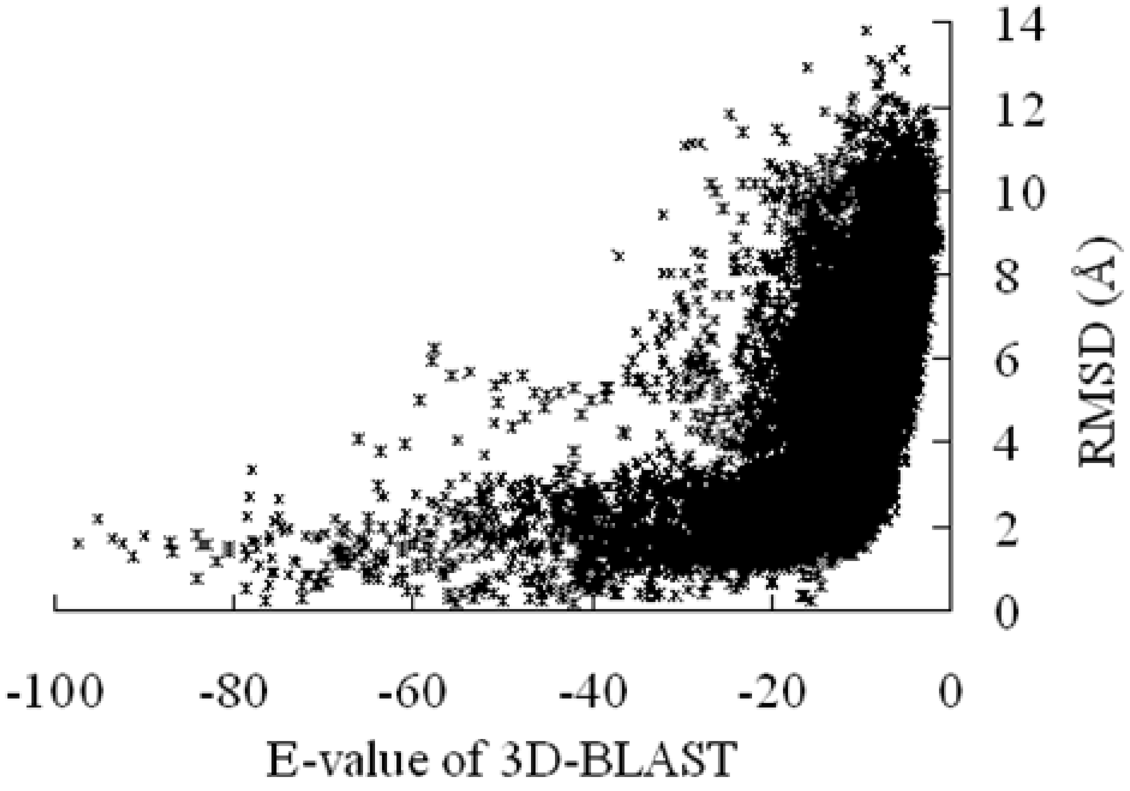


Figure 3.5 3D-BLAST performance with E values on the protein query set SCOP95-1.69

For classification assignment, the relationship between the *E*-value of the first hit and the number of correct (dark line) and false (gray line) classification assignments for the SCOP95-1.69 dataset were calculated (Figure 3.6). If the *E*-value was restricted to under e^{-15}, then 97.67% of 516 query structures are assigned correct classifications and the coverage was 91.47%. The coverage is defined as *P/T*, where *P* is the number of assigned structures by a method and *T* is total number of structures. For example, P is 472 and T is 516 for the set SCOP95-1.69. When the *E*-value was less than e^{-20} and e^{-10}, 99.31% and 95.26% of the predicted cases were correct, and the coverage values were 83.72% and 98.06%, respectively. When the sequence identity was less than 25% (154 proteins from among 516 proteins), the rate of correct assignment was 90.35%. The coverage was 72.12% when the *E*-value was less than e^{-15}. For the database search, the precision was 0.80 and the recall was 0.48 when the *E*-value was below e^{-15}; by comparison, the precision was 0.90 and the recall was 0.42 when the *E*-value was below e^{-20}. These analytical results demonstrate that the *E*-value of 3D-BLAST enables users to examine the reliability of the structure database search of a query.

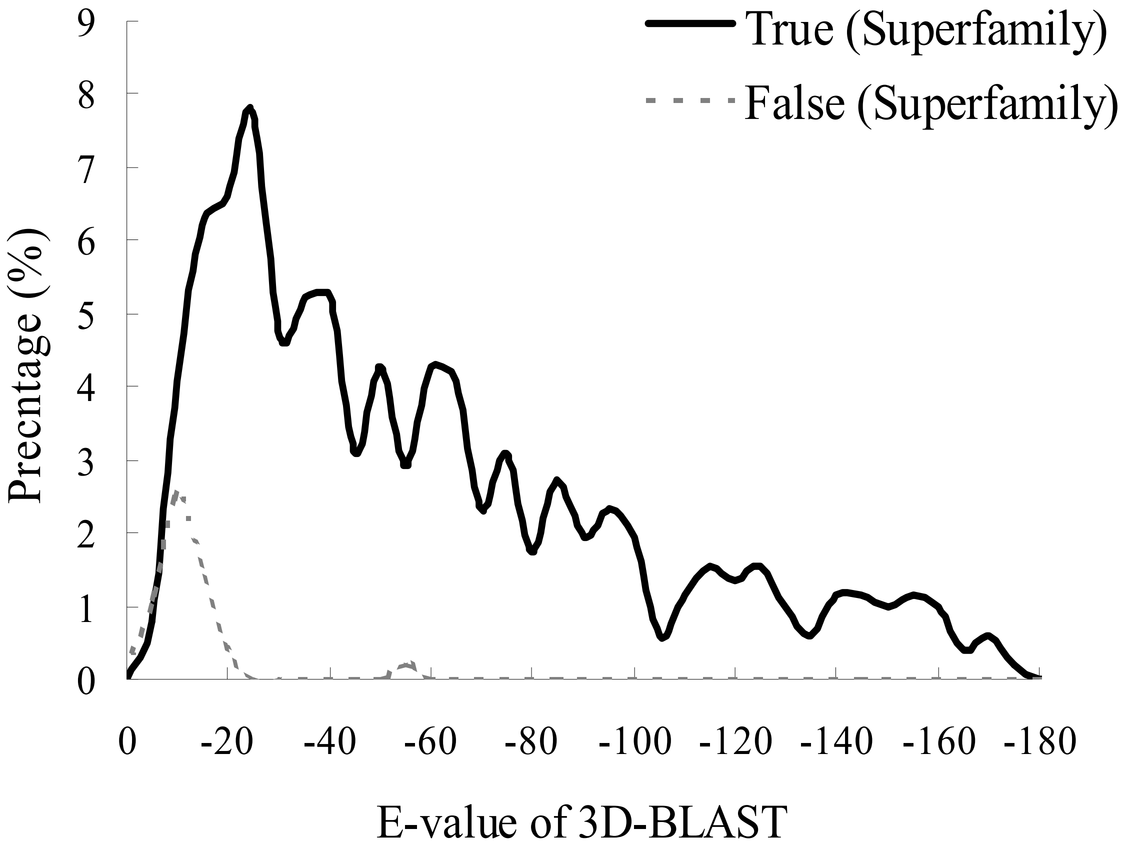


Figure 3.6 The relationship between *E*-values and the percentages of true and false superfamily assignment on the query set SCOP95-1.69.

3.5 Evolutionary Classification

3.5.1 3D-BLAST Database Search Examples

For many query proteins in SCOP-894, 3D-BLAST automatically recognizes the distantly related protein family members that escape standard sequence database similarity searches. Here, we discuss two examples involving protein families that have relatively weak sequence similarities. Tables 3.2 and 3.3 demonstrate these two cases. The first target is aminoglycoside N-acetyltransferase (NAT) AAC(6')-Iy [57] (PDB code 1s3z) (Figures 3.7 and 3.8). The secondary target is a structural genomics target (PDB code 1xi3) that is a member of a TIM beta/alpha-barrel fold [58] (Figure 3.9). In each case, 3D-BLAST reported a structurally and functionally relevant relationship in greater detail.

Table 3.2 3D-BLAST search results using aminoglycoside 6'-N-acetyltransferase as the query

PDB code	Protein name	SCOP family name	log(*E*-value)	rmsd (Å)	Sequence identity [a]	Species
1tiqA	Protease synthase and sporulation negative regulatory protein PaiA	N-acetyl transferase	-36.70	1.97	17	*Bacillus subtilis*
1qstA	GCN5 histone acetyltransferase	N-acetyl transferase	-32.70	3	14.4	*Tetrahymena thermophila*
1i12A	Glucosamine-phosphate N-acetyltransferase GNA1	N-acetyl transferase	-32.40	2.09	21.2	*Saccharomyces cerevisiae*
1gheA	Tabtoxin resistance protein	N-acetyl transferase	-29.70	2.36	21.5	*Pseudomonas syringae*
1qsoA	Histone acetyltransferase HPA2	N-acetyl transferase	-29.15	1.77	18.1	*Saccharomyces cerevisiae*
1cm0A	Histone acetyltransferase domain of P300/CBP associating factor	N-acetyl transferase	-29.05	2.8	16.4	*Homo sapiens*
1ufhA	Putative acetyltransferase YycN	N-acetyl transferase	-27.52	3.39	21.6	*Bacillus subtilis*
1vhsA	Putative phosphinothricin acetyltransferase YwnH	N-acetyl transferase	-26.40	2.68	18.3	*Bacillus subtilis*
1n71A	Aminoglycoside 6'-N-acetyltransferase	N-acetyl transferase	-26.40	2.28	18.8	*Enterococcus faecium*
1m44A	Aminoglycoside 2'-N-acetyltransferase	N-acetyl transferase	-25.52	2.96	18.9	*Mycobacterium tuberculosis*
1mk4A [b]	Hypothetical protein YqiY	N-acetyl transferase	-25.00	2.74	24.9	*Bacillus subtilis*
1p0hA [b]	Mycothiol synthase MshD	N-acetyl transferase	-24.30	1.51	14.2	*Mycobacterium tuberculosis*
1cjwA	Serotonin N-acetyltranferase	N-acetyl transferase	-24.22	3.04	16.6	*Ovis aries*
1bo4A [c]	Aminoglycoside 3-N-acetyltransferase	N-acetyl transferase	**-24.22**	**2.74**	**16.8**	*Serratia marcescens*
1nslA	Probable acetyltransferase YdaF	N-acetyl transferase	-23.52	2.92	18.1	*Bacillus subtilis*
1sqhA	Hypothetical protein cg14615-pa	Hypothetical protein cg14615-pa	**-21.00**	**2.39**	**15.7**	*Drosophila melanogaster*
1yghA	GCN5 histone acetyltransferase	N-acetyl transferase	-20.22	3.06	17.5	*Saccharomyces cerevisiae*
1q2yA	Probable acetyltransferase YjcF	N-acetyl transferase	-19.70	2.48	19	*Bacillus subtilis*
1bob	Histone acetyltransferase HAT1	N-acetyl transferase	-16.15	2.18	14.9	*Saccharomyces cerevisiae*
1ne9A2	Peptidyltransferase FemX	FemXAB	**-16.05**	**2.42**	**15.3**	*Weissella viridescens*
1lrzA3	Methicillin resistance protein FemA	FemXAB	-16.00	2.23	14.9	*Staphylococcus aureus*
1iicA1	N-myristoyl transferase	N-myristoyl transferase	-16.00	2.71	16.2	*Saccharomyces cerevisiae*
1iykA2	N-myristoyl transferase	N-myristoyl transferase	**-15.00**	**3.04**	**15.3**	*Candida albicans*
1fy7A	Histone acetyltransferase ESA1	N-acetyl transferase	-14.00	2.97	16.2	*Saccharomyces cerevisiae*
1ro5A	Autoinducer synthesis protein LasI	Autoinducer synthetase	**-13.22**	**3.37**	**19.2**	*Pseudomonas aeruginosa*
1iicA2	N-myristoyl transferase	N-myristoyl transferase	-13.10	2.61	16.8	*Saccharomyces cerevisiae*
1kzfA	Acyl-homoserinelactone synthase EsaI	Autoinducer synthetase	-12.70	3.74	13.7	*Pantoea stewartii subsp. Stewartii*
1iykA1	N-myristoyl transferase	N-myristoyl transferase	-12.30	2.85	18.6	*Candida albicans*
1lrzA2	Methicillin resistance protein FemA	FemXAB	-11.52	3.46	16.7	*Staphylococcus aureus*

[a] Sequence identity was calculated by FASTA software.

[b] These two proteins were found by PSI-BLAST if the threshold of the *E*-value was 0.01.

[c] The protein (bold case) is shown in Figure 6A.

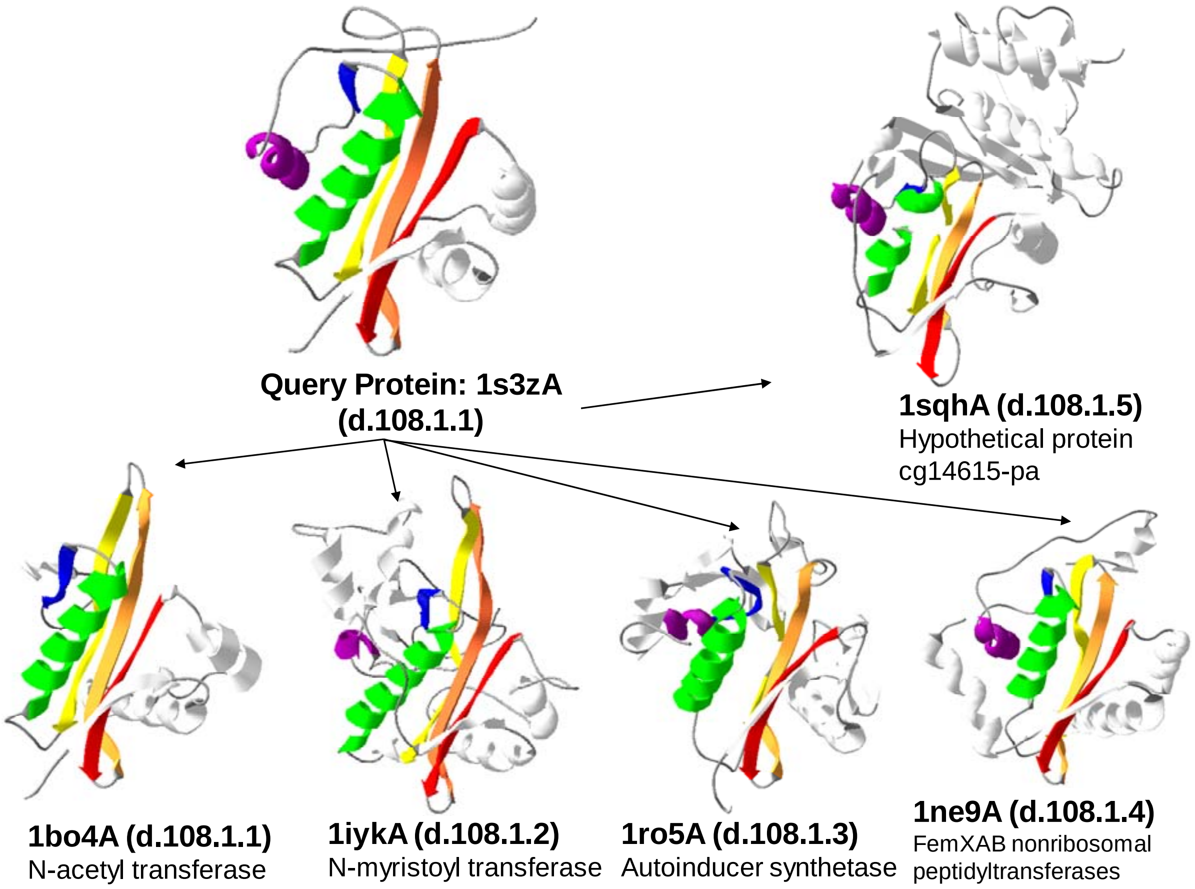


Figure 3.7 The structural recurrences of five homologous proteins from the NAT superfamily.

3.5.1.a N-acetyltransferases

The Salmonella enteritidis aminoglycoside N-acetyltransferase AAC(6')-Iy (PDB code 1s3z) is a member of the GCN5-related N-acetyltransferase (GNAT) superfamily [59] and the SCOP NAT superfamily. AAC(6')-Iy catalyzes acetyl group addition to aminoglycoside antibiotics, which are important antibacterial agents, and inhibits protein synthesis by inhibiting initiation and causing code misreading. Three conserved sequence motifs, termed D, A, and B, are characteristic of the GNAT superfamily, and motif A often contains a Arg/Gln-X-X-Gly-X-Gly/Ala motif (X denotes some variation) for the NAT family (Figure 3.8) [59].

3D-BLAST structural alphabet sequences

```
                     10        20        30        40        50        60        70        80        90       100       110       120
                      |         |         |         |         |         |         |         |         |         |         |         |
SS structure SSSSSS       SSSSSSS  SSSS                          SSS  SSSSSSS           HHHHHHHHHHHHHH         SSSSS         HHHHHHHH
Motif        DDDDDDDDDDDDDDDDDDDDDDDDDD                          AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA       BBBBBBBBBBBBBBBBBBBBBBBBBBBBB
             +++++     +  +++++++                                  +  +++++++   ++++   +M+++++++++++++   ++++++++++++     Q++++++++++++     E-value
    1s3zA    EFFFHH--V-SQ-PHVPFEH--EFNHXSTK----------VQP---M-T-MPFNF--HVPHEKKC-DQMQS---QMBACYBBYYCYACD---BSRNMPHHFHKK----CQPMBDYYBABBSRTKH
    1ufhA    EHFFNH--V-TLDNEXNEEH--KHEE--------------GQP---D-S-RHF-F--HVRXKKKC-BQXQS---RMGCBBBDCBAACDB---BSRHXPHHFHKF----BQNMDCDCYYCBSRNKH  -27.52
    1vhsA    HFFEFK--I-SR-TNVPFE----FHKFVT------------QP---D-D-QMQRK--EEEEEKKD-BRMQS---QMBYYYCACYBAGGB---DSRHMPEFFHKK----BQNMACACABBBSQNKN  -26.40
    1bo4A    EEEFEE--V-WR-TNMPEEHKFEFKEGLT------------QP--------FFEF--HXPNHKKC-DQMQS---QMCDYYBYBBBACDA---ASRNMPEEKH                         -24.22
    1sqhA    NEEFHKFDL-SN-NNXPEEK--------TK-----------IS---Q-T-RZ-------PNNKKI-DQMQS---QMACDYYAYCACACY---ALTRTNNHFHKK----CQNMDCYACBBBSRNKX  -21.00
    1ne9A2   HEEFHK--V-WR-THXPEFH--EF-HXW----------------------IP-NF--HZPEEFKV-WRT-T---QMGBBYBACYBBDAC---ASNNMPHETXPFVTQPLQPG-ACAADGDLR---  -16.05
    1iykA2   EEEFHF--KISR-NHXPEEH--FFHHFFFE----------MQT---GLN-MPFHEFKHXPEEK---GXMLTD--QMCACBYCADDBDCBGGGDSPEMPH--NKK-----QXWQMGLQGACSRHKN  -15.00
    1ro5A    HEEEHH--V-SQ-THXPEFE--EKFGDSPKCDLLLRMGGSTRTKKHX-TCGPF-E--HXPHHKHXTIRNMSKDLDDBDCYBAABYYAYB---BSRHXPEHKXNK----I---DDDBACYBSRTRH  -13.22
```

Amino acid sequences

```
                     10        20        30        40        50        60        70        80        90       100       110       120
                      |         |         |         |         |         |         |         |         |         |         |         |
    1s3zA    ASFIAM--A-DG-VAIGFAD--ASIRHDYV----------NGC---D-S-SPVVF--LEGIFVLP-SFRQR---GVAKQLIAAVQRWGT---NKGCREMASDTS----PENTISQKVHQALGFEE
    1ufhA    HLWSLK--L-NEKDIVGWLW--IHAE--------------PEH---P-Q-QEA-F--IYDFGLYE-PYRGK---GYAKQALAALDQAAR---SMGIRKLSLHVF----AHNQTARKLYEQTGFQE
    1vhsA    LYVAED--E-NG-NVAAWI----SFETFY------------GR---P-A-YNKTA--EVSIYIDE-ACRGK---GVGSYLLQEALRIAP---NLGIRSLXAFIF----GHNKPSLKLFEKHGFAE
    1bo4A    IALAAF--D-QE-AVVGALAAYVLPKFEQ------------PR--------SEIY--IYDLAVSG-EHRRQ---GIATALINLLKHEAN---ALGAYVIYVQ
    1sqhA    KSLGICRSD-TG-ELIAWIF--------QN-----------DF---S-G-LG-------XLQVLP-KAERR---GLGGLLAAAXSREIA---RGEEITLTAWIV----ATNWRSEALLKRIGYQK
    1ne9A2   RIFVAE--R-EG-KLLSTGI--AL-KYG----------------------RK-IW--YMYAGSMD-GNT-Y---YAPYAVQSEMIQWAL---DTNTDLYDLGGIESESTDDS-LYVFKHVFV---
    1iykA2   KSYVVE--DENG-IITDYFS--YYLLPFTV----------LDN---AQH-DELGIAYLFYYAS---DSFEKP--NYKKRLNELITDALITSKKFGVDVF--NCL-----TCQDNTYFLKDCKFGS
    1ro5A    YYMLIQ--E-DG-QVFGCWR--ILDTTGPYMLKNTFPELLHGKEAPC-SPHIW-E--LSRFAINSGQKGSLGFSDCTLEAMRALARYSL---QNDIQTLVTVTT----V---GVEKMMIRAGLDV
```

Figure 3.8 Sequence alignments of both structural alphabets and amino acid sequences of eight proteins from the NAT superfamily.

Using S. enteritidis AAC(6')-Iy as the query protein and an *E*-value cutoff of 10^{-10}, a 3D-BLAST search of the database SCOP1.69 found 19 members of the NAT family and 10 distantly related homologs of the NAT superfamily (Table 3.2). The sequence identities between the query protein and most of the homologous structures (25 of 29 proteins) were <20%. These 29 homologous proteins comprised 14 species. In contrast, a PSI-BLAST search of SCOP1.69 revealed only two hits (PDB code 1mk4A and 1pohA) with an *E*-value <0.01 in the NAT family (Table 3.2).

Figure 3.7 shows the structures of five distantly related proteins selected from different families of the NAT superfamily. These five proteins are N-acetyl transferase (PDB code 1bo4A), N-myristoyl transferase (PDB code 1iykA), autoinducer synthetase (PDB code 1ro5A), FemXAB nonribosomal peptidyltransferase (PDB code 1ne9A), and hypothetical protein cg14615-pa (PDB code 1sghA). The aligned structures are very similar, implying structural recurrence among these homologs. Each protein chain is drawn as a continuous-color spectrum from red through orange, yellow, green and blue to violet. Hence the N and C termini are red and violet, respectively. Table 3.2 shows the protein names, SCOP family names, the *E*-values, rmsd values, and sequence identities between these proteins and the query protein.

We produced both multiple structural letter sequence alignments and protein sequence alignments of eight proteins (Figure 3.8) using a simple star alignment method. This method uses the query protein as the center protein and seven-pair alignments between the query protein and seven hit homologous proteins. These eight proteins consisted of the six proteins shown in Figure 3.7 and two proteins (PDB code 1uth and 1vhs) selected from the NAT family. The alignments yielded several interesting observations, as follows. (1) For four NAT family proteins (PDB code 1s3zA, 1uthA, 1vhsA, and 1bo4A), 3D-BLAST automatically detected the invariant pattern (Arg88, Gln89, Arg90, Gly91, Val92, and Ala93 in the query protein) of motif A, which is responsible for the binding activity of the NAT family (red columns in Figure 3.8). (2) The 3D-BLAST structural alphabet sequences are much more conserved than amino acid sequences and this is the main reason that PSI-BLAST was unable to detect the invariant residues or to find these distantly related proteins. (3) The 3D-BLAST structural alphabet is also highly conserved in three motif areas (i.e., D, A, and B) of the NAT superfamily and in areas of secondary structures (i.e., S and H). (4) For these paired proteins, the structural alphabet sequence similarities correlate strongly with the *E*-values. These results demonstrate that 3D-BLAST can yield considerable information by unifying distantly related protein families into structurally and functionally conserved superfamilies.

3.5.1.b TIM Barrel Proteins

Thiamine phosphate pyrophosphorylase (PDB code 1xi3), an α/β protein with a triosephosphate isomerase (TIM) barrel fold [58], catalyzes the formation of thiamine phosphate—an essential nutrient for humans [60]. This protein is a structural genomics target for Southeast Collaboratory for Structural Genomics, which is a part of the Protein Structure Initiative [61]. The Pyrococcus furiosus enzyme was used as the query for a search of the SCOP 1.69 protein structure database. The TIM barrel has an eight-stranded ┐/β barrel and is by far the most common tertiary fold observed in protein crystal structures. Members of the TIM barrel family catalyze very different reactions and are attractive targets for protein engineering. Moreover, the ancestry of this enzyme remains unknown since there is limited sequence homology between TIM barrel proteins.

A

3D-BLAST structural alphabet sequences

```
                       10        20        30
                        |         |         |
                β7            α7        β8  α8
SS Structure   SSSSS        HHHHHH     SSS HHHH
Consensus       +++  ++ ++++ +++++S++++++H+ ++   E-value
  1xi3A   151  TEFHZVQN-TILQMYYYDGSRTVTFHHMILLA
  2tpsA   173  PEFHZVRH-NILQQGBACBSRTVTHHHMDLID  -67.70
  1qpoA   128  FEFHZDRH-KIDQGCACDGSRTMPKEHQGWLL  -27.40
  1w0mA1  175  FHEHXVQNXTCBYYCBDBGSQTVTFHHMBLLD  -27.40
  1vlwA   154  FEHEXDQH-TILQMBBABBSPNVTFHHVILL-  -27.52
```

Amino acid sequences

```
                       10        20        30
                        |         |         |
Consensus           +G    ++       G  +    ++
  1xi3A   151  PVVAIGGI-NKDNAREVLKTGVDGIAVISAVM
  2tpsA   173  PIVGIGGI-TIDNAAPVIQAGADGVSMISAIS
  1qpoA   128  MLESSGGL-SLQTAATYAETGVDYLAVGALTH
  1w0mA1  175  SVITGAGIESGDDVAAALRLGTRGVLLASAAV
  1vlwA   154  KFVPTGGV-NLDNVCEWFKAGVLAVGVGSAL-
```

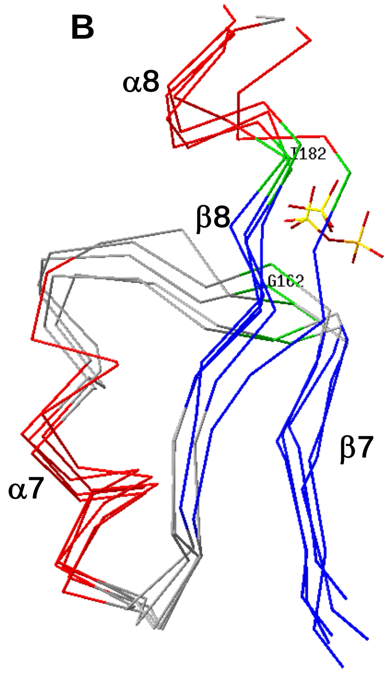


Figure 3.9 Multiple sequence alignments and multiple structure alignments resulting from a 3D-BLAST search using thiamine phosphate pyrophosphorylase from *Pyrococcus furiosus* as the query.

When the *E*-value was restricted to 10^{-15}, 3D-BLAST identified 74 members from 16 SCOP superfamilies containing a TIM barrel fold (Table 3.3). Figure 3.9 shows multiple sequence alignments and structure alignments of five homologous proteins derived from the 3D-BLAST pairing alignments. These proteins, thiamine phosphate synthase (PDB code 1xi3A and 2tpsA), quinolinic acid phosphoribosyltransferase (PDB code 1qpo), TIM (PDB code 1w0m), and aldolase (PDB code 1vlw), were selected from three different superfamilies. 3D-BLAST aligned the common phosphate-binding resides, ranging from β-7, loop-7, α-7, β-8 to α-8, on the last two loops of the barrel sheet [62] of these proteins. The secondary structures are indicated in red (helices) and blue (strands) and the loops are in gray. The phosphate-binding residues are indicated in green. Again, the structural alphabet sequences are highly conserved in this phosphate-binding site and are more conserved than amino acid sequences.

Table 3.3 Structure database search results of 3D-BLAST for finding homologous superfamilies using thiamine phosphate pyrophosphorylase from *Pyrococcus furiosus* as the query

SCOP superfamily	3D-BLAST [a]			
	Number of yielded proteins	Average log(*E*-value)	Average rmsd (Å)	Average sequence identity (%) [b]
Thiamin phosphate synthase	2	-98.3	0.71	66.2
Triosephosphate isomerase (TIM)	2	-25.0	2.41	22.9
Inosine monophosphate dehydrogenase	4	-23.3	2.89	18.8
Quinolinic acid phosphoribosyltransferase, C-terminal domain	2	-22.7	2.28	22.9
Phosphoenolpyruvate/pyruvate domain	6	-22.1	3.23	19.4
ThiG-like (Pfam 05690)	1	-22.0	2.95	23.4
RuBisCo, C-terminal domain	6	-21.9	2.76	17.9
Ribulose-phoshate binding barrel	19	-20.2	2.68	22.8
Aldolase	16	-18.7	2.79	21.1
UROD/MetE-like	1	-17.7	3.30	16.8
GlpP-like	1	-17.7	2.49	21.6
FMN-linked oxidoreductases	7	-17.6	2.82	18.2
Dihydropteroate synthetase-like	4	-16.8	2.74	21.0
Cobalamin(vitamin B12)-dependent enzymes	1	-16.7	2.99	15.0
CutC-like (Pfam 03932)	1	-16.4	2.46	19.4
Trans-glycosidases	1	-15.7	3.35	19.6

[a] Thresholds of the *E*-values was 10^{-15}.

[b] Sequence identity was calculated by FASTA.

3D-BLAST and PSI-BLAST produced 19 and 6 hits, respectively, for members of the ribulose-phosphate-binding barrel superfamily. The alignment results of both tools are similar, and the phosphate-binding residues are equivalently aligned (Figure 3.9). Because both alignment methods yielded confident hits, the homology between thiamine phosphate synthase and the ribulose-phosphate-binding barrel superfamily are considered reliable, despite the limited sequence identity. 3D-BLAST and PSI-BLAST also yielded similar alignments for other paired superfamilies: inosine monophosphate dehydrogenase and thiamine phosphate synthase, and FMN-linked oxidoreductases and thiamine phosphate synthase. These four SCOP superfamilies may be considered a homologous superfamily, termed the FMN-dependent oxidoreductase and phosphate-binding enzymes (FMOP) family, as proposed by Nagano et al. [58]

3D-BLAST identified five homologous superfamilies, including quinolinic acid phosphoribosyltransferase, phosphoenolpyruvate, and dihydropteroate synthetas*E*-like. These distant relationships were also reported by Nagano et al. [58] using PSI-BLAST or IMPALA [63] with different iteration numbers. In addition, 3D-BLAST and sequential structure alignment program (SSAP) [64] yielded two distantly related superfamilies (RuBisCo and trans-glycosidases), but PSI-BLAST or IMPALA could not find these two relationships. However, SSAP was unable to identify two superfamilies (triosephosphate isomerase and dihydropteroate synthetase-like) that could be retrieved by 3D-BLAST, PSI-BLAST and IMPALA. The above observations suggest that 3D-BLAST may be able to identify new links between SCOP superfamilies.

3.5.1.c Yeast copper chaperone for superoxide dismutase

Using the yeast copper chaperone for superoxide dismutase (yCCS) from Arabidopsis thaliana (PDB code 1JK9-B) [65] as the query protein and an *E*-value threshold of 10^{-10}, a 3D-BLAST search of the database SCOP1.69 found 19 members (Table 3.4). Figure 3.10 shows two hits of the search results. The protein (yCCS) comprised amino-terminal and carboxyl-terminal domains. The amino-terminal domain, called HMA (heavy-metal associated) domain in the SCOP database, plays a role in copper delivery. This domain contains an MH/TCXXC metal binding motif (blue box in Figure 3.10A), and is very similar to the metallochaperone protein Atx1. The carboxyl-terminal domain, termed the Cu,Zn superoxide dis-mutase-like domain in the SCOP database, comprised an eight-stranded β-barrel that strongly resembles yeast superoxide dismutase I and human superoxide dismutase I.

Table 3.4 3D-BLAST search results by copper chaperone for superoxide dismutase (PDB code 1JK9-B) from yeast as query

PDB code	Protein title	log(E-value)	rmsd (Å)	Sequence identity (%) [a]	SCOP sccs	*Species*
1EJ8-A	Copper chaperone for yeast sod	-50.70	1.10	57.6	b.1.8.1	*Saccharomyces cerevisiae*
1QUP-A	Copper chaperone for superoxide dismutase	-27.05	0.58	28.3	d.58.17.1	*Saccharomyces cerevisiae*
1CC8-A	Superoxide dismutase 1 copper chaperone	-17.40	1.64	8.6	d.58.17.1	*Saccharomyces cerevisiae*
1TO4-A	Superoxide dismutase	-17.22	2.78	19.6	b.1.8.1	*Schistosoma mansoni*
1DO5-A	Human copper chaperone for superoxide dismutase domain II	-16.30	2.57	17.3	b.1.8.1	*Homo sapiens*
1OSD-A	Oxidized Merp from Ralstonia metallidurans CH34	-16.05	1.61	11.1	d.58.17.1	*Ralstonia metallidurans*
1Q0E-A	Copper, Zinc Superoxide Dismutase	-14.22	1.68	17.7	b.1.8.1	*Bos taurus*
1OAL-A	Superoxide dismutase	-14.00	2.19	17.7	b.1.8.1	*Photobacterium leiognathi*
1SRD-A	Copper, Zinc Superoxide Dismutase	-13.30	2.71	17.5	b.1.8.1	*Synthetic construct*
1FE0-A	Copper transport protein atox1	-13.10	1.40	9.9	d.58.17.1	*Homo sapiens*
1OZU-A	Copper, Zinc Superoxide Dismutase	-12.70	2.42	18.5	b.1.8.1	*Homo sapiens*
1ESO	Copper, Zinc Superoxide Dismutase	-12.30	2.49	17.6	b.1.8.1	*Escherichia coli*
1FVQ-A	Copper-transporting Atpase	-12.00	1.64	9.9	d.58.17.1	*Saccharomyces cerevisiae*
1JCV	Copper, Zinc Superoxide Dismutase	-11.70	2.24	20.3	b.1.8.1	*Saccharomyces cerevisiae*
1S6U-A	Copper-transporting ATPase 1	-11.15	1.87	8.6	d.58.17.1	*Homo sapiens*
1XSO-A	Copper, Zinc Superoxide Dismutase	-10.70	1.88	19.3	b.1.8.1	*Xenopus laevis*
1OQ3-A	Potential copper-transporting ATPase	-10.40	1.84	11.4	d.58.17.1	*Bacillus subtilis*
1VCA-A	Human vascular cell adhesion molecule-1	-10.30	3.76	15.9	b.1.1.3	*Homo sapiens*
1KQK-A	Potential copper-transporting ATPase	-10.22	1.63	12.3	d.58.17.1	*Bacillus subtilis*
1MWY-A	The N-terminal domain of ZntA in the apo-form	-10.10	1.67	9.0	d.58.17.1	*Escherichia coli*

[a] Amino acid sequence identity is calculated using FASTA software. PDB, Protein Data Bank; rmsd, root mean square deviation.

A (N-terminal domain)

```
Structural alphabet sequence:  Identities = 29/68 (42%)
1jk9B:   2 HKHFEFHKNXXSLQGCBYYAB-DQMQPGQHVTEFEEMLSRTFEEXTQTFCBAAABABBBSQPEEFHVP 68
1cc8A:   1 HFHEEFHKHHVTIGDCBACBDCDGGQXCQPVPEEEEGLSRPEEEXPQTFBDYYYYYCGGSQTHE--VP 66

Amino acid sequence:  Identities = 12/68 (17%)
1jk9B:   2 TYEATYAIPMHCENCVNDIKA-CLKNVPGINSLNFDIEQQIMSVESSVAPSTIINTLRNCGKDAIIRG 68
1cc8A:   1 IKHYQFNVVMTCSGCSGAVNKVLTKLEPDVSKIDISLEKQLVDVYTTLPYDFILEKIKKTGKEV--RS 66
```

B (C-terminal domain)

```
Structural alphabet sequence:  Identities = 49/160 (30%)
1jk9B:  75 PFEFFHKFMTNKCQTFDQTKQHXNEEEHFHXTLLNFEEEHEHXPNMSXVPFFHKNKGPRHXVSVLGGSKT--HVPFFXPFK---------EXXTKFDWTDQPEHEEFHEHNQPKD-DCQ--VQTFHE----HHKNFMPQ---MQPMSRHXPNHHEKNPKN 213
1q0eA:   1 PEEFFHHNDS----------QHNKHEEHFH-TWLPFENZNEEXRTN-VVNKNNNNKMPRHXVSVLMQSZTTRNMSXNNRKNIQPGQTKCQHTXNFE-FKISRNEFHXHHHMGPKMVWIQNXVQTFHEKKVSNEKTQXSQTCGGDLTTQNXTHVPEKKHKT 147

Amino acid sequence:  Identities = 23/160 (14%)
1jk9B:  75 SAVAILETFQKYTIDQKKDTAVRGLARIVQVGENKTLFDITVNGVPEAGNYHASIHEKGDVSKGVESTGK--VWHKFDEPI---------ECFNESDLGKNLYSGKTFLSAPLPT-WQL--IGRSFV----ISKSLNHP---ENEPSSVKDYSFLGVIAR 213
1q0eA:   1 KAVCVLKGDG----------PVQGTIHFEA-KGDTVVVTGSITGLT-EGDHGFHVHQFGDNTQGCTSAGPHFNPLSKKHGGPKDEERHVGDLGNVT-ADKNGVAIVDIVDPLISLSGEYSIIGRTMVVHEKPDDLGRGGNEESTKTGNAGSRLACGVIGI 147
```

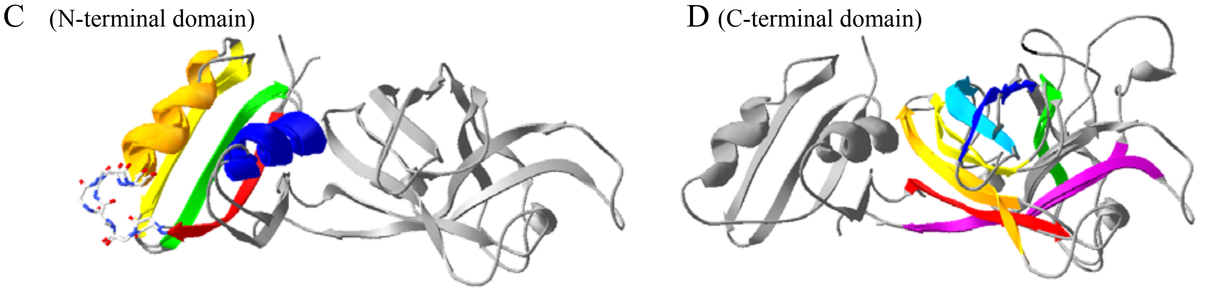


Figure 3.10 Sequence and structure alignments of 3D-BLAST search results using yCCS as the query.

3D-BLAST was able to identify 9 and 10 homologous structures of amino-terminal domains and carboxyl-terminal domains, respectively, using this two-domain protein (yCCS) as query. The sequence identities between yCCS and most of the homologous structures (17 out of 19 proteins) were less than 20%. Figures 3.10A and 3.10C illustrate sequence alignments and the structure alignment between yCCS and an amino-terminal domain homologous protein (PDB code 1CC8-A [66]). The sequence identities of structure alphabet and amino acid sequences were 42% and 17%, respectively. 3D-BLAST can align six amino acids of the metal binding motif together, and the rmsd is 1.64 Å between these two proteins. The aligned secondary structures are represented as a continuous color spectrum from red through orange, yellow, green and blue to violet. Figures 3.10B and 3.10D show the sequence and structure alignments between yCCS and a carboxyl-terminal domain homologous protein (PDB code 1QO*E*-A [67]). The sequence identities of the structure alphabet and the amino acid sequences were 30% and 14%, respectively, and the rmsd between these two proteins was is 1.68 Å. The structural alphabets were strongly conserved in areas of the secondary structures (green block), which are β-strands represented by structural alphabets, such as E, F, H, K, and N. These results reveal that the structural alphabet sequences are much better conserved than the amino acid sequences, which explains why 3D-BLAST could detect the invariant residues and find these distantly related proteins.

3.5.2 Structural genomics targets

We analyzed 319 structural genomics targets, called SG-319, using 3D-BLAST with regard to function assignment. The structural genomics initiative aims to determine representative structures for all protein families in cells [1, 2, 68]. To sample the protein structural space more efficiently, structural genomics projects employ various target selection strategies to filter out proteins that are homologous to the proteins with structures already in the PDB [3]. As a result, the molecular functions of the proteins targeted by structural genomics are often unknown. The SG-319 set contains 319 structural genomics targets contributed by more than 10 structural genomics consortia, and publication dates range from 1 January 2005 to 30 September 2005. There are 126 proteins in SG-319 having the 'unknown function' annotation.

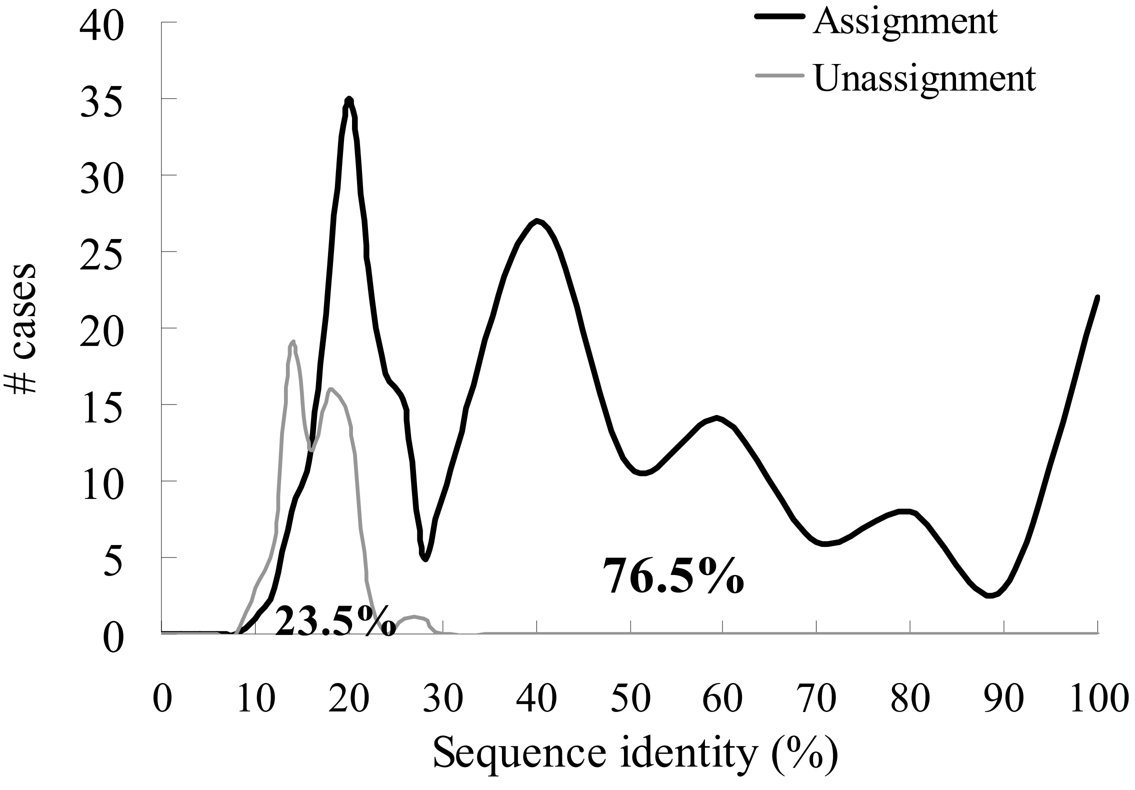


Figure 3.11 3D-BLAST function assignment results for 319 proteins targeted by structural genomics.

3D-BLAST used these 319 proteins as query proteins, and the search classification database was SCOP 1.69, which contains 12,074 domains. About 38.2% (122 proteins) and 32.6% (104 proteins) of the SG-319 proteins have more than 25% and under 20% sequence identity, respectively, to one of the library representatives of the SCOP superfamily, according to search results with 3D-BLAST. In all, 3D-BLAST assigned 244 (78.5%) proteins to SCOP superfamilies if the threshold of *E*-value was set at under e^{-15} by the SG-319 query set (Figure 3.11). When the sequence identity was more than 25%, 98.4% (120 out of 122) of these cases could be assigned to a SCOP superfamily by 3D-BLAST, and 62.9% (124 out of 197) of the remaining proteins could also be assigned.

The following observations help in comparing the characteristics and performance between applying 3D-BLAST to SG-319 (Figure 3.11) and applying it to SCOP95-1.69 (Figure 3.6). First, the distribution of the sequence identity of these two sets was significantly different. The sequence identities of 197 (61.8%) and 154 (29.85%) proteins in SG-319 and SCOP95-1.69, respectively, were under 25%. The average sequence identity in SG-319 is significantly lower than that of SCOP95-1.69. Second, the assigned parentages of SG-319 and SCOP95-1.69 were 78.5% and 91.47%, respectively, when the *E*-value was restricted to under e^{-15}. If the sequence identity was under 25%, then the assigned rates were 62.9% (SG-319) and 72.12% (SCOP95-1.69). Third, 3D-BLAST achieved similar accuracies for both sets if the sequence identity was above 25%. These observations are consistent with recent analyses of proteins targeted by structural genomics [3, 69].

Figure 3.12 shows that 3D-BLAST assigned a structural genomics target (PDB code 1YRH) to the flavodoxin-related family [70] based on the first-rank protein (PDB code 1E5D [71]) in the hits. The *E*-value was 10^{-25} and the Z score of CE and rmsd value were 5.7 and 1.56 Å, respectively, when these two proteins were aligned. These two proteins have the same Gene Ontology (GO) annotations [72] and the same domain annotations in three databases, including PROSITE [38, 73], Pfam [39], and CATH [56]. The aligned structures of these two proteins are similar, and the FMN-binding motifs (wireframe model) are also aligned well (Figure 3.12). Eight of the top 10 proteins in the hits are the members of the same SCOP superfamily. However, PSI-BLAST was unable to yield the same assignment.

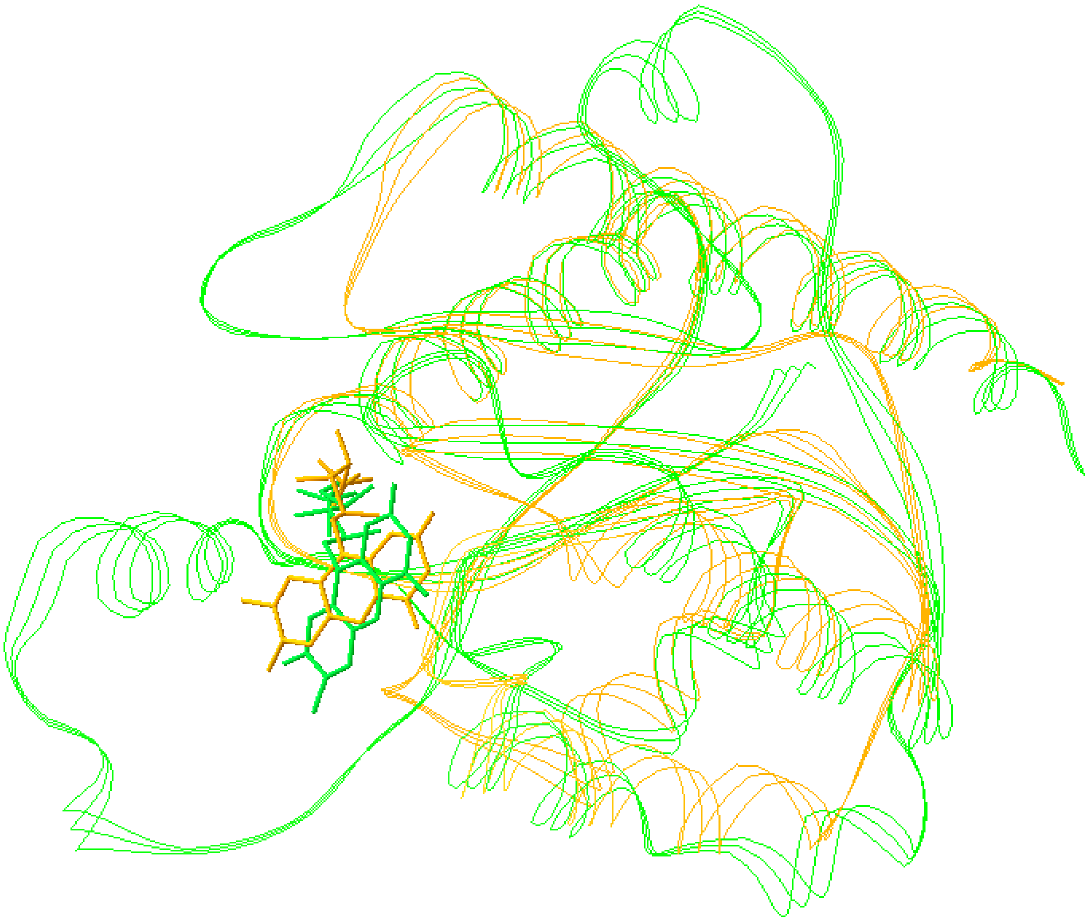


Figure 3.12 Structure alignment between the one of structural genomics target (1yrhA, green) and the first-rank protein (PDB code 1e5dA, orange) in the hit list

3.6 Method comparison

3.6.1 Comparison with PSI-BLAST

Table 3.5 shows the accuracies of 3D-BLAST and PSI-BLAST in structure database searches and evolutionary classification assignments using the query protein set SCOP-894. Here, we compare 3D-BLAST with PSI-BLAST because PSI-BLAST is often much better than BLAST for these purposes. We installed standalone PSI-BLAST [7] on a personal computer with a single processor (Pentium 2.8-GHz with 512 Mbytes). The search databases and substitution matrixes are the main differences between 3D-BLAST and PSI-BLAST. In 3D-BLAST, the substitution matrix is the SASM and the searching database is SADB; in contrast, PSI-BLAST uses an amino acid sequence database and the substitution matrix is BLOSUM62. The number of iterations for PSI-BLAST is set at 3. Since the gap penalty is an important factor, we systematically tested various combinations of gap penalty for

3D-BLAST and the SASM matrix. Here, the optimum values of the open gap penalty and the extended one are 8 and 2, respectively.

Table 3.5 Comparison of 3D-BLAST and PSI-BLAST for automatic SCOP structural function assignment on the protein query set SCOP-894

Class name	894 proteins			Sequence identity <25%		
		3D-BLAST	PSI-BLAST		3D-BLAST	PSI-BLAST
	Number of queries	Corrected assignment percentage	Corrected assignment percentage	Number of queries	Corrected assignment percentage	Corrected assignment percentage
All alpha	161	94.41%	94.41%	36	75.00%	66.67%
All-beta	199	94.47%	93.97%	49	77.55%	73.33%
α/β	292	97.26%	91.44%	66	87.88%	65.75%
α+β	242	94.63%	88.84%	78	83.33%	60.87%

SCOP-894, as shown Table 3.1.

For most sets of sequence identities, 3D-BLAST outperforms PSI-BLAST (Table 3.5). Nearly 74.4% (665 of 894) of query proteins are >25% identical to one of the library representatives from the same SCOP superfamily and ~99.5 % of these domains can be correctly mapped by both 3D-BLAST and PSI-BLAST. As expected, the accuracy of both methods is comparable for the 25% sequence identity cutoff. The accuracies are 95.8% (3D-BLAST) and 94.0% (PSI-BLAST) if the sequence identity ranges from 20% to 25%. When the sequence identity is <20% (122 of 894 proteins), the accuracy of 3D-BLAST ranges from 52.8% to 78.4%, whereas the accuracy of PSI-BLAST ranges from 21.6% to 46.9%. These proteins are more difficult to assign due to limited similarity of the query proteins to the representative library domains. In addition, the ROC curve provides an estimation of the likely number of true-positive and false-positive predictions for a database search tool. Based on ROC curves, 3D-BLAST is much better in this respect than PSI-BLAST.

3D-BLAST yields significantly better results than PSI-BLAST when working at sequence identity levels of 25%. One prevalent difficulty in making classification assignments by automatic methods is correctly assigning proteins that have very limited sequence similarity to the library representatives. Thus, the general observation is that, as expected, sequence comparison tools that are more sensitive to distant homology typically are more successful at making challenging assignments. These results show that 3D-BLAST achieves more reliable assignments than PSI-BLAST in cases of low sequence identity.

The false assignments made by 3D-BLAST (41 proteins) and by PSI-BLAST (73 proteins) were compared among 894 query proteins. Indeed, 28 query proteins were given false assignments by both 3D-BLAST and PSI-BLAST. Only 13 proteins were simultaneously given correct assignments by PSI-BLAST and false assignments by 3D-BLAST. Conversely, 45 proteins of the missed assignments made by PSI-BLAST were correctly mapped by 3D-BLAST. Most of the remaining proteins assigned by 3D-BLAST but not identified by PSI-BLAST represent cases that are typically difficult for sequence alignment methods. For the 41 assignments that 3D-BLAST missed, the sequence identity was <20% and the *E*-values of 9 cases were more than the threshold (i.e., e^{-15}). For 46% proteins of these 41 missed cases, the correct superfamily assignment can be determined using the top 5 ranked hits.

The factors causing 3D-BLAST to generate 41 false assignments can be roughly divided into five categories. The first factor is that the actual Euclidean distances were not considered in the structural alphabet. Therefore, 3D-BLAST may have made minor shifts when aligning two local segments with similar codes, such as segments a and a’ shown in Figure 3.1E. Therefore, 3D-BLAST is more sensitive when the query proteins are members of the “all alpha” (e.g., PDB code 1v2z [64] and 1owa [74]) or “all beta” (e.g., PDB code 1sq9 [75] and 1ri9 [76]) classes in SCOP. In the second category, the structural similarity of a query protein to the representative library domains is very limited (e.g., PDB code 1sp3 [77] and 1q5f [78]). In the third category, the query proteins had multiple domains (e.g., PDB code 1s35 [79] and 1tua). 3D-BLAST can correctly assign these two cases if domains are used as query targets. In the fourth category, an inherent problem of the BLAST algorithm is a lack of detecting remote homology of structural alphabet sequences. Use of PSI-BLAST as the search algorithm for 3D-BLAST slightly improved the overall performance on the set SCOP-894, and this procedure correctly assigned four cases (PDB code 1pa4 [80], 1sq9 [75], 1ovy [81], and 1t3k [82]) among these 41 false cases. An enhanced position-specific score matrix of the structure alphabet for SADB databases should be developed to improve the performance of 3D-BLAST. The final factor is that the *E*-values of the hits are not significant.

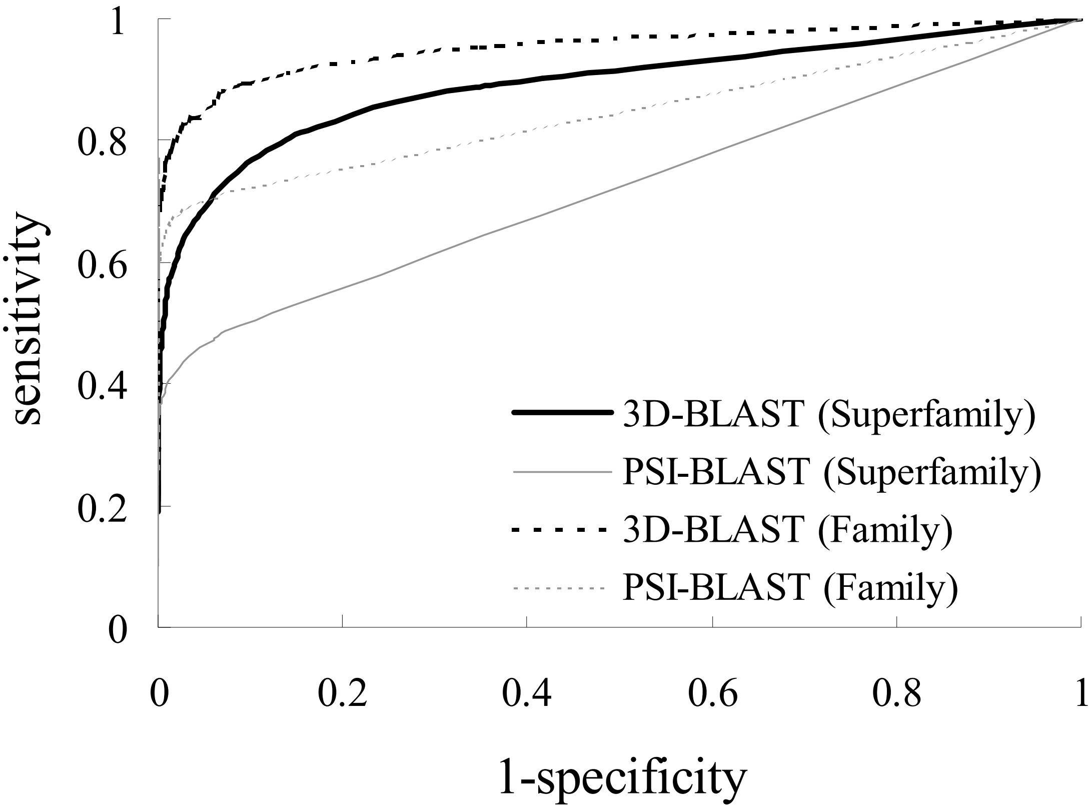


Figure 3.13 Evaluation of the 3D-BLAST and PSI-BLAST in database search based on ROC curves

Figures 3.13 and 3.14 illustrate the accuracies of the 3D-BLAST and PSI-BLAST in structure database searches and evolutionary classification assignments using the query set SCOP95-1.69. For this experiment, 3D-BLAST was compared with PSI-BLAST, because PSI-BLAST often performs much better than BLAST for this purpose. For a database search tool, the ROC curve (Figure 3.13) provides an estimation of the likely number of true positive and false positive predictions. A perfect method, which can recover all true hits without any false positives, can be denoted as a point in the top left corner of this graph, whereas a random method that generates equal numbers of true positive and false positive predictions uniformly distributed across all scores would yield a diagonal line from (0,0) to (1,1). Figure 3.13 shows that 3D-BLAST (dark lines) yields much better predictions than does PSI-BLAST (gray lines). The sensitivity of family assignments was superior to that of superfamily assignments in both methods, whereas the false-positive rates of family assignments were higher than those of the superfamily assignments.

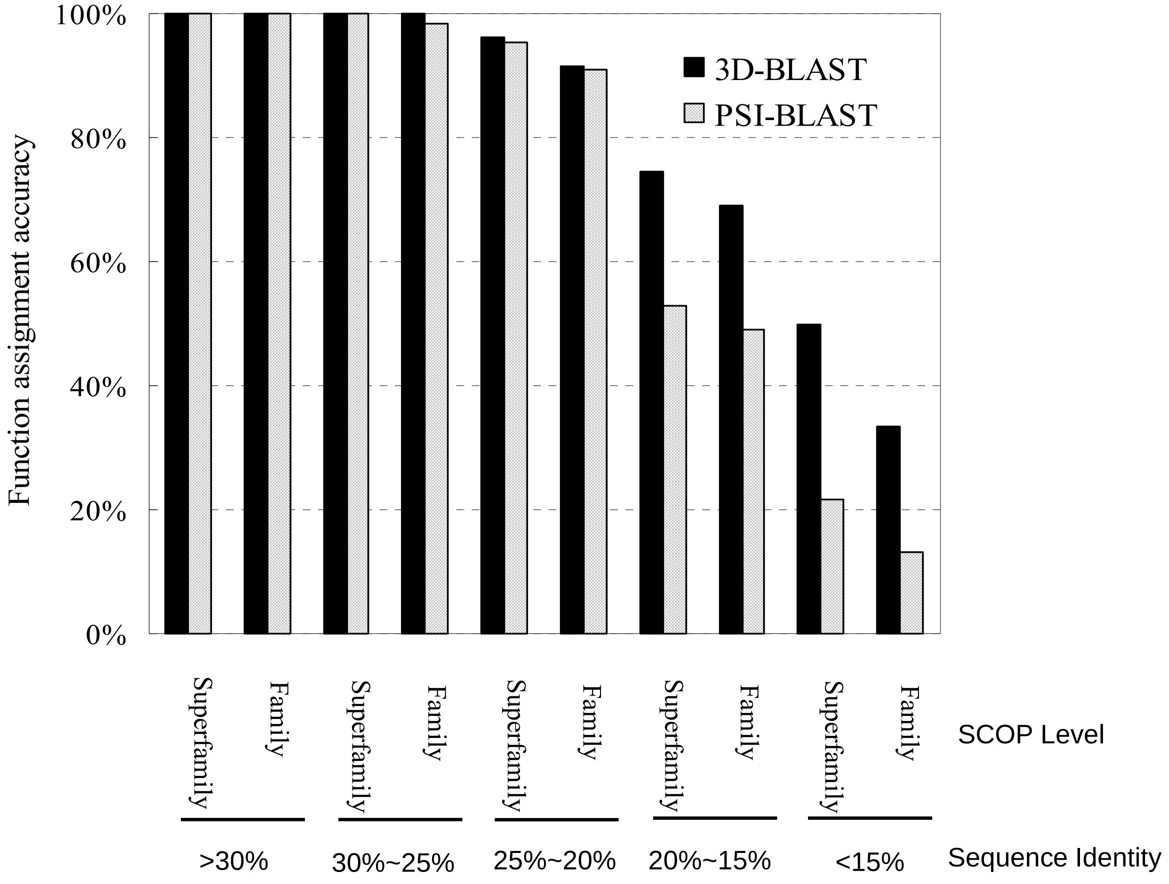


Figure 3.14 Comparison 3D-BLAST with PSI-BLAST: The percentages of correct classification assignments.

For most sets of sequence identities, 3D-BLAST outperformed PSI-BLAST (Figure 3.14) in protein evolutionary classification assignments. Almost 70.16% (362 out of 516 proteins) of query proteins were more than 25% identical to one of the library representatives from the same SCOP superfamily, and 100% of these domains were correctly mapped by both 3D-BLAST and PSI-BLAST. When the sequence identity was less than 25% (154 out of 516 proteins), the accuracy of 3D-BLAST ranged from 96.29% to 50%, whereas the accuracy of PSI-BLAST ranged from 94.29% to 21.74% (Figure 3.14). These proteins were difficult to assign because of the limited similarity of the query proteins to the representative library domains. 3D-BLAST yielded significantly better results than did PSI-BLAST at sequence identity levels of 25% or less. The analytical results reveal that, as expected, sequence comparison tools that are more sensitive to distant homology are usually more successful at making challenging assignments. In summary, 3D-BLAST achieved more reliable assignments than did PSI-BLAST in cases of low sequence identity for this test set. The structural alphabet, SADB database, and SASM matrix could predict protein structures more accurately than simple amino acid sequence analyses.

3.6.2 Comparison with others

Comparing the results of different structure database search methods is generally neither straightforward nor completely fair, because each such method utilizes different accuracy measures, searching databases, and test complexes. Figure 3.15 shows the relationship between recall and precision, and Table 3.6 presents the average search time and average precision of 3D-BLAST, PSI-BLAST, MAMMOTH, CE, TOPSCAN, and ProtDex2 on 108 query proteins proposed by Aung and Tan [12]. The performance of TOPSCAN and ProtDex2, which are fast search methods for scanning structure databases, was summarized from previous studies [12]. Other four programs were installed and run on the same personal computer with a single processor. Here, the PSI-BLAST and 3D-BLAST used *E*-values to order the hit proteins; MAMMOTH and CE (detailed structure alignment tools) utilized Z scores to rank the hit proteins.

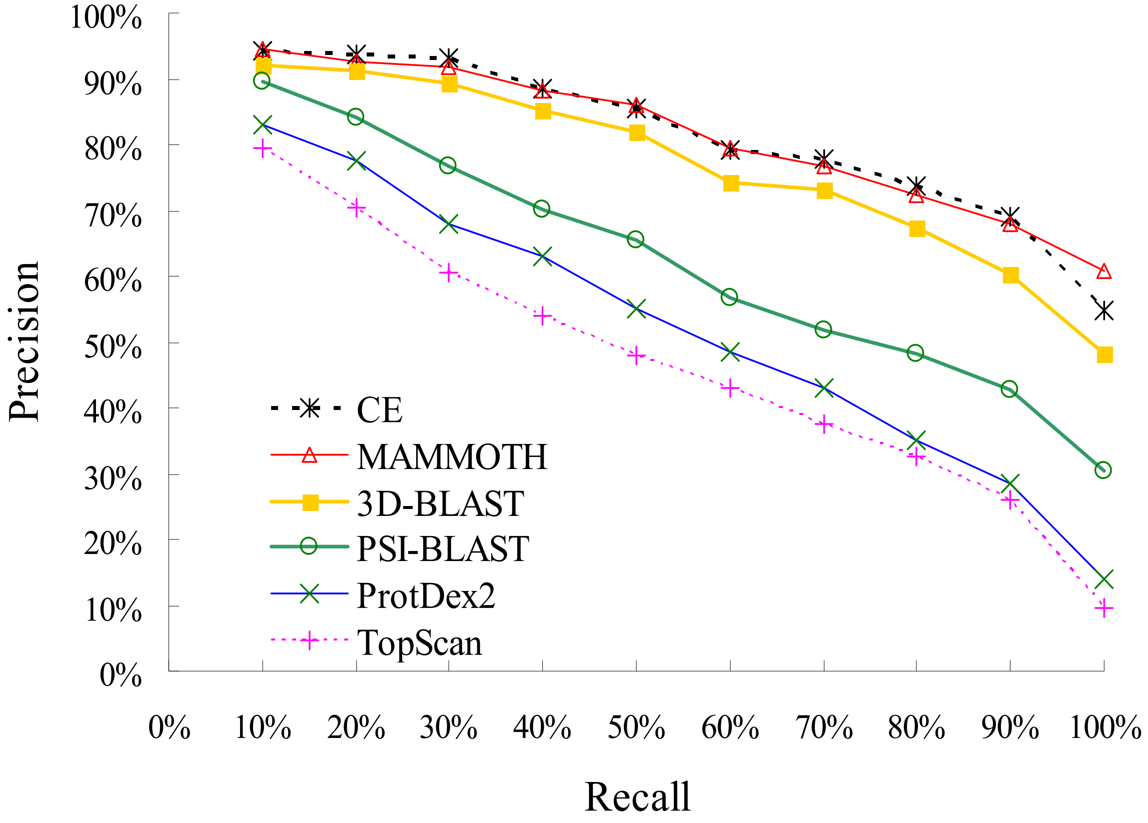


Figure 3.15 3D-BLAST versus fast structure search, sequence profile search, and detailed structural alignment.

Table 3.6 Average search time and mean average precision of each program on 108 queries in SCOP-108

Program	Mean of average precision	Total searching time (s)	Average time per query (s)	Related to 3D-BLAST
PSI-BLAST [a]	69.8%	18.31	0.170	0.533
3D-BLAST	78.2%	34.35	0.318	1
MAMMOTH [b]	82.1%	131,855	1220.88	3838.58
CE	83.4%	~13.5 days	~3 hours	~34000

[a] PSI-BLAST used *E*-values to rank the hit proteins

[b] MAMMOTH and CE utilized Z-scores to rank hit proteins.

Time was measured using a personal computer equipped with an Intel Pentium 2.8 GHz processor with 1024 Mbytes of RAM memory.

On average, 3D-BLAST required about 3.18 seconds to scan the database for each query protein (Table 3.6). It is about 34,000 and 3838 times faster than CE and MAMMOTH, respectively. 3D-BLAST was about two times slower than PSI-BLAST, because 3D-BLAST identified many more words (typically of length three for proteins in BLAST) that score more than a threshold value in the SADB databases than those identified by PSI-BLAST in protein sequence databases. The reason for this stems from the fact that the BLAST algorithm scans the database for words that score at least a threshold when aligned with some words within the query sequence; the algorithm then extends each such 'hit' in both directions to check the alignment score [7].

MAMMOTH is the best and TOPSCAN is the worst for these 108 queries among these six methods (Figure 3.15). 3D-BLAST was much better than fast structure database search methods (TOPSCAN and ProtDex2), and its performance approached those of CE and MAMMOTH. Notably, PSI-BLAST outperformed both TOPSCAN and ProtDex2, which considered secondary and 3D protein structures. As shown in Table 3.6, the mean of average precision of 3D-BLAST (78.2%) was better than that of PSI-BLAST (69.2%) and lightly worse than those of CE (82.1%) and MAMMOTH (83.4%). For some query proteins, such as serotonin N-acetyltranferase [83] (PDB code 1CJW-A) and translation initiation factor IF2/ eIF5B [84] (PDB code 1G7S-A), 3D-BLAST, MAMMOTH, and CE were markedly better than PSI-BLAST because most sequence identities between the query proteins and their members are under 20%. For several query proteins, such as human dihydro-orotate dehydrogenase [85] (PDB code 1D3G-A) and yeast copper chaperones for SOD [86] (PDB code 1EJ8-A), CE and MAMMOTH were worse than 3D-BLAST. Interestingly, PSI-BLAST outperformed CE, MAMMOTH, and 3D-BLAST for S-adenosylhomocysteine hydrolase [87] (PDB code 1B3R-A).

The recognition performance of 3D-BLAST is expressed as top rankings, using Lindahl's benchmark [88], together with the performance of eight popular sequence comparison (for example, HMM and profile methods). The benchmark includes 976 proteins derived from the SCOP for identifying homologous pairs at different similarity levels. Sequence identities between the query proteins and their homologous members in the superfamily and fold levels are much lower than those at the family level. These methods can be divided into two categories: methods using only single sequence information (BLAST2 and SSEARCH) and methods using multiple sequence alignments (PSI-BLAST, HMMER-HSSP [89], HMMER-PSI-BLAST [89], SAM-HSSP [55], SAM-PSI-BLAST [55], and BLAST-LINK [88]). The methods of constructing profiles/HMMs used a larger dataset, comprising the

SWISSPROT-35 and TREMBL-5 databases [90] together with the benchmark sequences of the HSSP database [91].

At the family level, 3D-BLAST identified 78.4% of homologous pairs that were ranked in the top 5. This was comparable to the best performance of any of the other methods (78.9%), which was achieved by BLAST-LINK. At the superfamily and fold levels, 3D-BLAST significantly outperformed all of the other methods. 3D-BLAST yielded 54.8% and 39.3% homologous pairs at the superfamily and fold levels, respectively. On the other hand, the best accuracies for the other methods were 40.6% (by BLAST-LINK) at the superfamily level and 18.7% (by SAM-PSI-BLAST) at the fold level.

Table 3.7 Average search time and performance of each program on 50 proteins selected from SCOP95-1.69

Program	Average time of a query (seconds)	Average time of a pair alignment (seconds)	Relative to 3D-BLAST	Correct assignment percentage	Mean of average precision
3D-BLAST	1.298	0.000118	1	94%	85.20%
PSI-BLAST	0.483	0.0000458	0.37	84%	68.16%
YAKUSA	8.880	0.0008072	6.84	90%	74.86%
MAMMOTH	1834.18	0.1667285	1413.08	100%	94.01%
CE	22053.32	2.0047	16990	98%	90.78%

Time was measured using a personal computer equipped with an Intel Pentium 2.8-GHz processor with 512 Mbytes of RAM memory. SCOP95-1.69 is described in Table 3.1.

Table 3.7 shows the average search time and average precision of 3D-BLAST, PSI-BLAST, YAKUSA, MAMMOTH, and CE on 50 query proteins. These five programs were installed and run on the same personal computer with a single processor. Here, the PSI-BLAST used *E*-values to order the hit proteins; YAKUSA, MAMMOTH, and CE utilized Z-scores to rank hit proteins. Because ~228 days are required to evaluate CE on each query in the set SCOP-894, we uniformly selected 50 proteins from the set SCOP95-1.69 based on the lengths of these 516 query proteins. On average, 3D-BLAST required ~1.298 seconds to scan the database for pattern hits for each query protein (this time included system overhead). 3D-BLAST is 16,990 and 1,413 times faster than CE and MAMMOTH, respectively. 3D-BLAST is lightly faster than YAKUSA and ~3 times slower than PSI-BLAST, which searches amino acid sequence databases. We found that 3D-BLAST was as fast as BLAST when their performance was similar. In our tests, 3D-BLAST was slightly slower than

BLAST because 3D-BLAST identified many more hit words in SADB databases compared with those identified by PSI-BLAST in protein sequence databases. The reason stems from the fact that the BLAST algorithm scans the database for hit words that score more than a threshold value when aligned with words in the query sequence; it then extends each hit word in both directions to check the alignment score.

Among these five methods, MAMMOTH is the best and PSI-BLAST is the worst for these 50 queries (Table 3.7). The means of average precision of 3D-BLAST (85.20%) was better than PSI-BLAST (68.16%) and YAKUSA (74.86%) as well as approached those of CE (90.8%) and MAMMOTH (94.01%). For some query proteins, such as Polyketide synthase associated protein 5 [92] (PDB code 1q9jA), Hypothetical protein Alr5027 (structural genomics target and PDB code 1vl7A), and avrpphf orf1 [93] (PDB code 1s28), 3D-BLAST, MAMMOTH, and CE were markedly better than PSI-BLAST because most sequence identities between the query proteins and their members are < 20%. For several query proteins, such as Calcium-dependent protein kinase sk5 [94] (PDB code 1s6iA) and Putative mar1 (structural genomics target and PDB code 1x9gA), CE was worse than 3D-BLAST because CE ranks some false positive proteins prior to ranking true positive cases. Interestingly, PSI-BLAST lightly outperformed CE and 3D-BLAST for GTP-binding protein YPT1 [95] (PDB code 1ukvY) and 1s6iA [94].

The main factors causing 3D-BLAST to perform poorly on some cases in both SCOP95-1.69 and SCOP-108 datasets are summarized as follows. First, 3D-BLAST might have made minor shifts when aligning two local segments with similar codes, because the structural alphabets did not consider the actual Euclidean distances. Hence, 3D-BLAST is more sensitive when the query proteins (for example, PDB code 1VDL-A and 1PMZ-A in SCOP95-1.69) are the members of the 'all-α' class in SCOP. Second, the structural similarity of a query protein to the library members is rather limited. Third, an inherent problem in the BLAST algorithm is inability to detect remote homology of structural alphabet sequences. Use of PSI-BLAST as the search algorithm for 3D-BLAST slightly improved the overall performance on the SCOP95-1.69 set. An enhanced position-specific score matrix of the structure alphabet for SADB databases should be developed to improve the performance of 3D-BLAST in the future. Finally, the *E*-values of the hits are not significant.

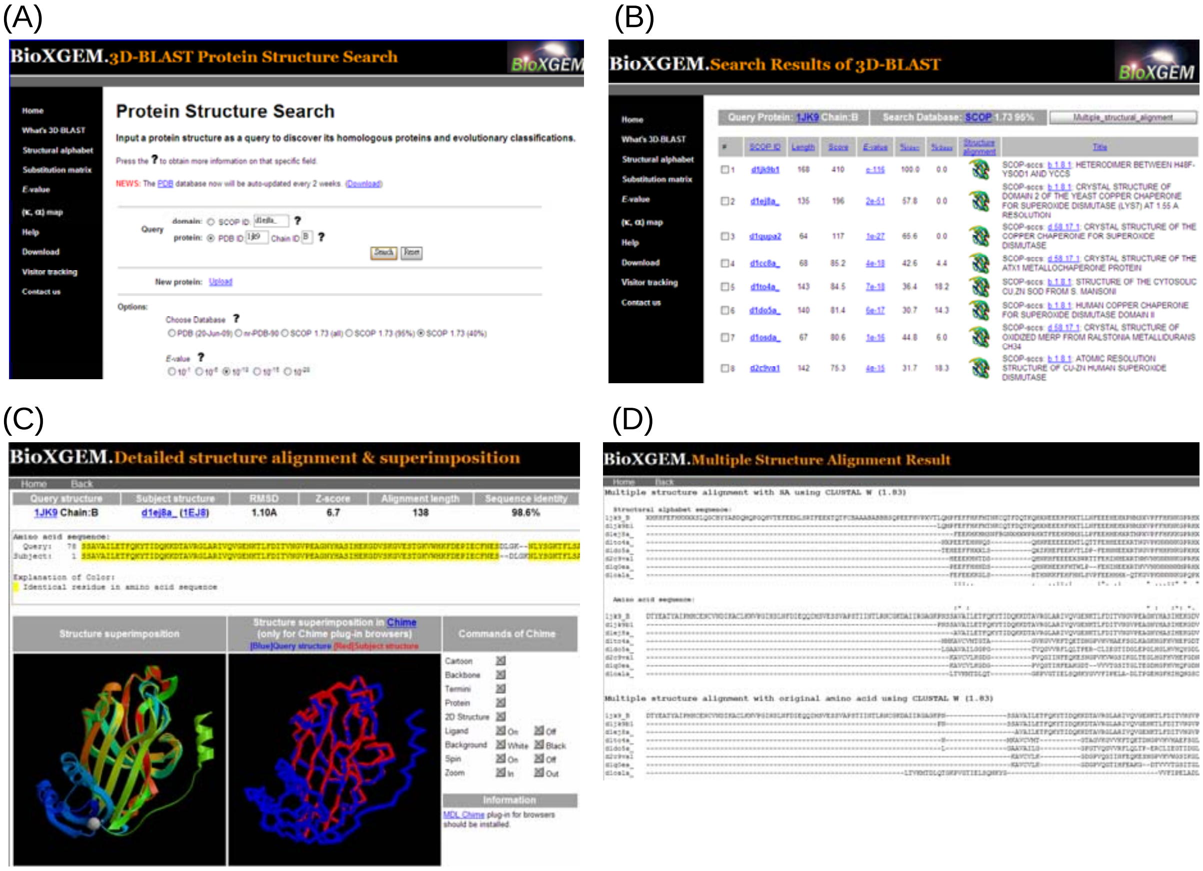
We demonstrated the robustness and adaptability of 3D-BLAST for the initial scan of large protein structure databases; conversely, detailed structure alignment tools often align two structures slowly but accurately. Because of basic differences, comparisons between 3D-BLAST and detailed structure alignment tools are not straightforward. However, detailed structure alignment tools can be applied to refine the searching structures of 3D-BLAST to improve accuracy of prediction.

3.7 Web service

We have built a website server for global users (http://3d-blast.life.nctu.edu.tw/). Figure 3.16 shows the interface and procedures of 3D-BLAST search. Users may input a PDB code with a protein chain (e.g. 1JK9-B) or a SCOP identifier (Figures 3.16A). Otherwise, users need to upload a three-dimensional protein domain structure with PDB format. And then, users need to choose which SADB will be searched. There are 5 kinds of available structural databases, including PDB, nr-PDB-90, SCOPall, SCOP95%, and SCOP40%. These SADB will be checked from the latest version of the source database and auto-updated every 2 weeks. Besides, users may select the options of *E*-value. This setting is a threshold for reporting matching protein structures against structural database. The *E*-value means that such matches are expected to be found merely by chance. The lower *E*-value is more stringent, causing to fewer number of matches being reported. Users are able to see more detail information about explanation of 3D-BLAST input on website (http://3d-blast.life.nctu.edu.tw/help.php).

After users click the search button, our server search and discover the structural homologous of query protein structure in few seconds (Figure 3.16B). The web page of search results can present the description of each SA sequence alignment among query and its homologous proteins, including the title of homologous protein, alignment length, *E*-value, the percent of structural alphabet identity and gaps.

In search results page, there are hyper-links which can link and execute detailed structure alignment using CE tool for structure superimposition between query and subject structures. Figure 3.16C shows that the aligned structures are visualized not only in PNG format using MolScript and Raster3D packages but also in 3D model with Chime software. Our server allows users to download the aligned structure coordinates in PDB format. Besides, 3D-BLAST server also provide both multiple sequence alignments and multiple structural alignments (Figure 3.16D) based on users' requirements in search results page. The server uses ClustalW software to multiple align structural alphabet and amino acid sequences of various proteins respectively. Additionally, the number of global queries of web service is more than 10,000 from June 2006 to June 2009.



BioXGEM.3D-BLAST Protein Structure Search

Protein Structure Search

Input a protein structure as a query to discover its homologous proteins and evolutionary classifications.

Press the ? to obtain more information on that specific field.

NEWS: The PDB database now will be auto-updated every 2 weeks. (Download)

Query domain: ○ SCOP ID d1ej8a_ ?
protein: ● PDB ID 1jk9 Chain ID B ?
Search Reset

New protein: Upload

Options:
Choose Database ?
○ PDB (20-Jun-09) ○ nr-PDB-90 ○ SCOP 1.73 (all) ○ SCOP 1.73 (95%) ● SCOP 1.73 (40%)

E-value ?
○ 10^{1} ○ 10^{6} ● 10^{12} ○ 10^{15} ○ 10^{20}

BioXGEM.Search Results of 3D-BLAST

Query Protein: 1JK9 Chain:B | Search Database: SCOP 1.73 95% | Multiple_structural_alignment

#	SCOP ID	Length	Score	E-value	%iden	%gap	Structure alignment	Title
□1	d1jk9b1	168	410	e-115	100.0	0.0		SCOP-sccs: b.1.8.1: HETERODIMER BETWEEN H48F-YSOD1 AND YCCS
□2	d1ej8a_	135	196	2e-51	57.8	0.0		SCOP-sccs: b.1.8.1: CRYSTAL STRUCTURE OF DOMAIN 2 OF THE YEAST COPPER CHAPERONE FOR SUPEROXIDE DISMUTASE (LYS7) AT 1.55 A RESOLUTION
□3	d1qupa2	64	117	1e-27	65.6	0.0		SCOP-sccs: d.58.17.1: CRYSTAL STRUCTURE OF THE COPPER CHAPERONE FOR SUPEROXIDE DISMUTASE
□4	d1cc8a_	68	85.2	4e-18	42.6	4.4		SCOP-sccs: d.58.17.1: CRYSTAL STRUCTURE OF THE ATX1 METALLOCHAPERONE PROTEIN
□5	d1to4a_	143	84.5	7e-18	36.4	18.2		SCOP-sccs: b.1.8.1: STRUCTURE OF THE CYTOSOLIC CU,ZN SOD FROM S. MANSONI
□6	d1do5a_	140	81.4	6e-17	30.7	14.3		SCOP-sccs: b.1.8.1: HUMAN COPPER CHAPERONE FOR SUPEROXIDE DISMUTASE DOMAIN II
□7	d1osda_	67	80.6	1e-16	44.8	6.0		SCOP-sccs: d.58.17.1: CRYSTAL STRUCTURE OF OXIDIZED MERP FROM RALSTONIA METALLIDURANS CH34
□8	d2c9va1	142	75.3	4e-15	31.7	18.3		SCOP-sccs: b.1.8.1: ATOMIC RESOLUTION STRUCTURE OF CU-ZN HUMAN SUPEROXIDE DISMUTASE

BioXGEM.Detailed structure alignment & superimposition

Home Back

Query structure	Subject structure	RMSD	Z-score	Alignment length	Sequence identity
1JK9 Chain:B	d1ej8a_ (1EJ8)	1.10A	6.7	138	98.6%

```
Amino acid sequence:
  Query:   78 SSAVAILETFQKYTIDQKKDTAVRGLARIVQVGENKTLFDITVNGVPEAGNYHASIHEKGDVSKGVESTGKVWHKFDEPIECFNESDLGK--NLYSGKTFLSA
Subject:    1 SSAVAILETFQKYTIDQKKDTAVRGLARIVQVGENKTLFDITVNGVPEAGNYHASIHEKGDVSKGVESTGKVWHKFDEPIECFNES--DLGKNLYSGKTFLSA

Explanation of Color:
  Identical residue in amino acid sequence
```

Structure superimposition | Structure superimposition in Chime (only for Chime plug-in browsers) [Blue]Query structure [Red]Subject structure | Commands of Chime

Cartoon ☒
Backbone ☒
Termini ☒
Protein ☒
2D Structure ☒
Ligand ☒On ☒Off
Background ☒White ☒Black
Spin ☒On ☒Off
Zoom ☒In ☒Out

Information
MDL Chime plug-in for browsers should be installed.

BioXGEM.Multiple Structure Alignment Result

Home Back

Multiple structure alignment with SA using CLUSTAL W (1.83)

Structural alphabet sequence:

Amino acid sequence:

Multiple structure alignment with original amino acid using CLUSTAL W (1.83)

Figure 3.16 The interface and procedures of 3D-BLAST web service.

3.8 Summary

As more protein structures become available and structural genomics efforts provide structural models in a genome-wide strategy, there is a growing need for fast and accurate methods for discovering homologous proteins and evolutionary classifications of newly determined structures. We have developed 3D-BLAST, in part, to address these issues. 3D-BLAST is as fast as BLAST and calculates the statistical significance (*E*-value) of an alignment to indicate the reliability of the prediction. Using this method, we first identified 23 states of the structural alphabet that represent pattern profiles of the backbone fragments and then used them to represent protein structure databases as structural alphabet sequence databases (SADB). Our method enhanced BLAST as a search method, using a new structural alphabet substitution matrix to find the longest common substructures with high-scoring structured segment pairs from an SADB database. Using personal computers with Intel Pentium4 (2.8 GHz) processors, our method searched more than 10,000 protein structures in

1.3 seconds and achieved a good agreement with search results from detailed structure alignment methods.

Chapter 4

Recognizing Protein Structural Domains and SCOP Superfamilies

4.1 Introduction

As protein structures become increasingly available and structural genomics provide structural models in a genome-wide strategy [1], proteins with unassigned functions are accumulating and the number of protein structures in the Protein Data Bank (PDB) is rapidly rising [4]. The evolutionary classification databases, such as SCOP [43, 96] and CATH [56], are valuable resources for understanding protein functions, structural similarity and evolutionary relationships. However, these two widely used databases are updated intermittently using manual and semi-automated methods. This current structure-function gap clearly reveals the need for powerful automated methods to classify protein domains based on their tertiary structures and is important in producing manually tuned classification databases.

Many automatic domain classification approaches have been developed to determine homologs and evolutionary classifications [97, 98] of a query structure. Protein sequence database search tools, such as BLAST, PSI-BLAST and Superfamily [98], are useful computational tools. However, these tools are commonly unreliable in detecting remote homologous relationships that are indicated by such structural alignment tools as DALI, MAMMOTH and SSM [99]. Structural alignment tools typically take several seconds to align two known structures. At this speed, about one day is required to compare a single protein structure with those in PDB. SCOPmap [97], which is computationally more expensive, combines sequence and structural information for SCOP superfamily assignment.

Recently, we have proposed a fast and efficient tool, called 3D-BLAST [34, 35], to quickly search similar structures. This tool is as fast as BLAST and provides the statistical significance (*E*-value) of an alignment to indicate the reliability of a structure. 3D-BLAST outperformed fast structural search methods (TOPSCAN and YAKUSA) and approached the performance of detailed structural alignment approaches (CE and MAMMOTH). 3D-BLAST is rapid and accurate in scanning a large protein structural database, and is useful in an initial scan for similar protein structures, which can be refined using detailed structural comparison methods. However, several factors that deteriorate 3D-BLAST's performance are (a) 3D-BLAST may have made minor shifts in aligning two local segments with similar letters, because the structural alphabet do not consider actual Euclidean distances, (b) the E-values of the hit proteins are insignificant, and (c) the query is a multiple-domain protein.

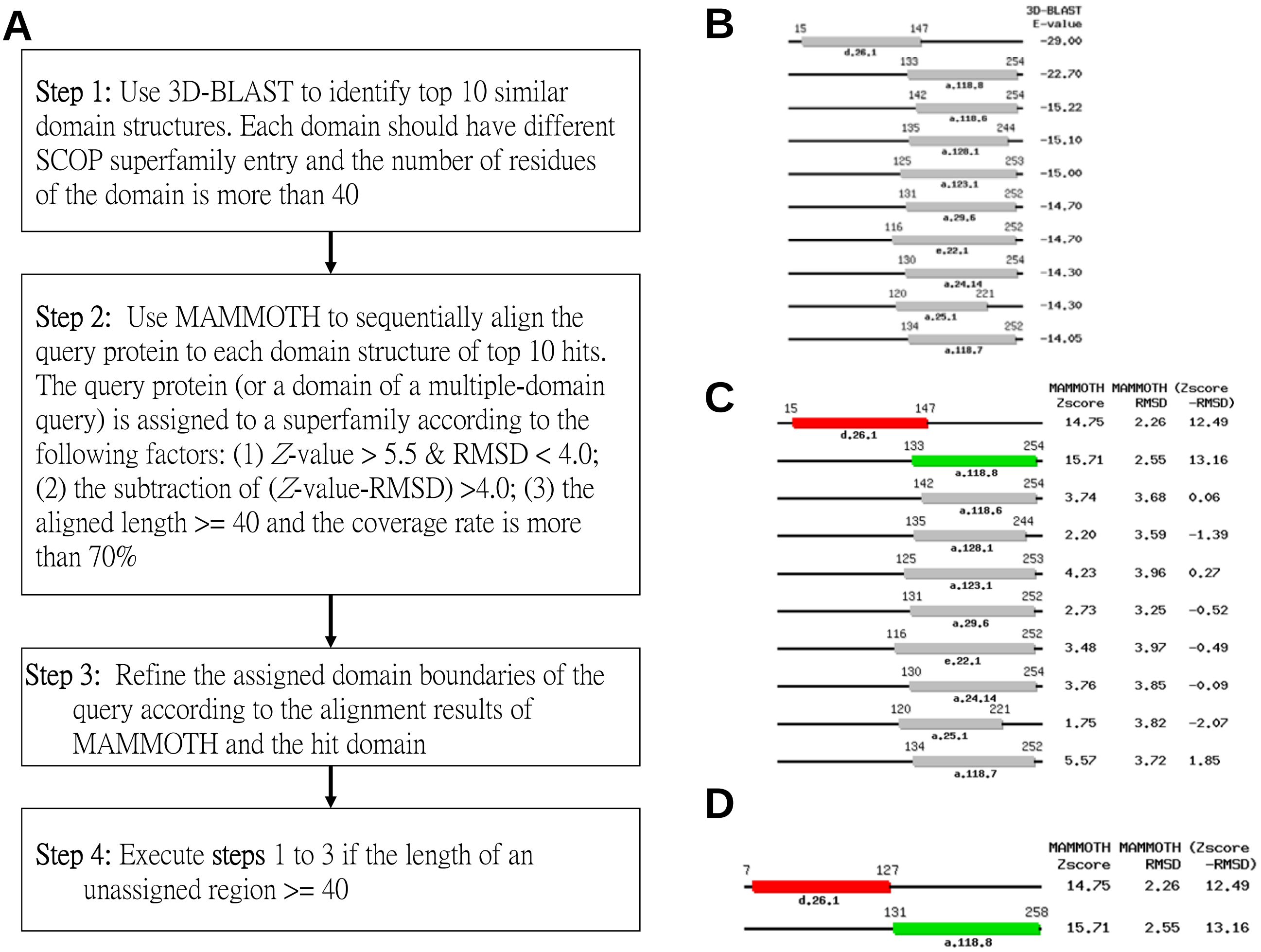
This work presents an automated server (fastSCOP), which integrates a fast structure database search tool (3D-BLAST) and a detailed structural alignment tool (MAMMOTH), to recognize SCOP domains and evolutionary superfamilies of a query structure. The classification accuracy of this server is 98% for 464 single-domain queries and 122 multiple-domain queries. After a query structure is assigned to a superfamily, this server is able to provide both multiple sequence alignments and multiple structural alignments of the selected members in a SCOP superfamily.

4.2 Materials and methods

Figure 4.1 presents an overview of the fastSCOP server for rapidly recognizing SCOP domains and evolutionary superfamilies. This sever uses 3D-BLAST to scan quickly the SCOP 1.71 database and selected the top ten hit domain structures, which are associated with different SCOP superfamily entries (Figure 4.1B). MAMMOTH was then adopted to align sequentially the query structure with each structure of the top ten structures, to refine the domain boundaries and to recognize SCOP superfamilies (Figures 4.1C and 4.1D). Our previous work [34, 35] demonstrated that 3D-BLAST required ~1.4 seconds to scan the structural domains in SCOP 1.69 and was 16,990 and 1,413 times faster than CE and MAMMOTH, respectively. These two detailed structural alignment tools perform similarly on the test set; MAMMOTH was ~12 times faster than CE. The SCOP 1.71 database (October 2006) has 75,930 domains that are derived from 27,599 PDB entries (Jan 18, 2005). The numbers of folds, superfamilies and families are 971, 1,589 and 3,004, respectively. 3D-BLAST requires structural alphabet sequence databases (SADB) for fast scanning a protein structural database. In this work, we created an SADB derived from known domain structures (12,927 domains) in SCOP1.71 with <95% identity to each other based on the (κ, α) plot [34, 35].

The fastSCOP server performs four main steps to identify the SCOP domains and superfamilies. First, 3D-BLAST was adopted to identify the similar structures (hit SCOP domains), which are ordered by E-value, of a query structure from an SADB database (Figure

4.1B). 3D-BLAST is the first tool to provide fast search of a protein structural database using the BLAST, which searches on a SADB database with a structural alphabet substitution matrix (SASM) [34, 35]. The fastSCOP then selected the top ten hit domains that have different SCOP superfamily entries. Based on the structural alphabet alignments between the query and hit SCOP domains, this sever can identify multiple domains if a multiple-domain structure is queried. For each hit domain, the aligned length should be more than 40 residues and the coverage rate of two neighbor hit domains should be less than 10%.



Step 1: Use 3D-BLAST to identify top 10 similar domain structures. Each domain should have different SCOP superfamily entry and the number of residues of the domain is more than 40

Step 2: Use MAMMOTH to sequentially align the query protein to each domain structure of top 10 hits. The query protein (or a domain of a multiple-domain query) is assigned to a superfamily according to the following factors: (1) *Z*-value > 5.5 & RMSD < 4.0; (2) the subtraction of (*Z*-value-RMSD) >4.0; (3) the aligned length >= 40 and the coverage rate is more than 70%

Step 3: Refine the assigned domain boundaries of the query according to the alignment results of MAMMOTH and the hit domain

Step 4: Execute steps 1 to 3 if the length of an unassigned region >= 40

A
B
3D-BLAST
E-value
15 147 d.26.1 -29.00
133 254 a.118.8 -22.70
142 254 a.118.6 -15.22
135 244 a.128.1 -15.10
125 253 a.123.1 -15.00
131 252 a.29.6 -14.70
116 252 e.22.1 -14.70
130 254 a.24.14 -14.30
120 221 a.25.1 -14.30
134 252 a.118.7 -14.05
C
MAMMOTH Zscore MAMMOTH RMSD (Zscore -RMSD)
15 147 d.26.1 14.75 2.26 12.49
133 254 a.118.8 15.71 2.55 13.16
142 254 a.118.6 3.74 3.68 0.06
135 244 a.128.1 2.20 3.59 -1.39
125 253 a.123.1 4.23 3.96 0.27
131 252 a.29.6 2.73 3.25 -0.52
116 252 e.22.1 3.48 3.97 -0.49
130 254 a.24.14 3.76 3.85 -0.09
120 221 a.25.1 1.75 3.82 -2.07
134 252 a.118.7 5.57 3.72 1.85
D
MAMMOTH Zscore MAMMOTH RMSD (Zscore -RMSD)
7 127 d.26.1 14.75 2.26 12.49
131 258 a.118.8 15.71 2.55 13.16

Figure 4.1 Overview of the fastSCOP server for SCOP domain recognition and superfamily assignment.

After the top ten hit SCOP domains were identified, this server applied MAMMOTH to align sequentially the query structure with each structure of these hit domains, ordered by E-value. For each structural alignment, MAMMOTH yielded the Z-score and root-mean-square deviation (RMSD) of the C_α atom positions of the aligned residues between the query structure and the hit structure (Figure 4.1C). The query structure (or one domain of a multiple-domain protein) was assigned to a SCOP superfamily when the pair-structure alignment satisfied the following criteria: (a) the Z-score exceeds 5.5; (b) the RMSD value is less than 4 Å; (c) the subtraction value, Z-score-RMSD, exceeds 4.0; and (d) the number of the aligned residues exceeds 40 and the coverage rate between the query protein (domain) and hit domain exceeds 75%. In the third step, the fastSCOP refined the boundaries (the start and end positions) of the assigned domain according to the aligned regions and the sequence length of the hit domain (Figure 4.1D). Finally, the fastSCOP executed steps 1 to 3 when the length of the unassigned region of the query structure was more than 40 residues.

4.3 Experimental Results and Discussion

4.3.1 Results

A query protein set, SCOP-586 (Table 4.1), was selected to evaluate the utility of the fastSCOP server for recognizing the structural domains and evolutionary superfamilies of a query structure. The SCOP-586 query set has 464 single-domain proteins and 122 multiple-domain proteins that are in SCOP 1.69 but not in SCOP 1.67, and the search database was SCOP 1.67 (11,001 structures). Among the 122 multiple-domain queries, 104 proteins have two domains, 14 have three domains and 4 have more than four domains. The total number of domains is 272 in the multiple-domain query set and the total number of domains in the SCOP-586 is 736.

Table 4.1 presents the accuracy of superfamily assignment and the average execution time of the fastSCOP, 3D-BLAST and MAMMOTH on the query set SCOP-586. Standalone fastSCOP, 3D-BLAST and MAMMOTH were run on a personal computer with a single Pentium 2.8 GHz processor with 1024 Mbytes RAM. The 3D-BLAST and MAMMOTH used E-values and Z-scores, respectively, to order the hit proteins. For 3D-BLAST, the top rank of a hit list of a query was selected as the SCOP superfamily. For MAMMOTH, the same criteria (Z-score>5.5; RMSD value<4 Å and (Z-score-RMSD)>4.0) of the fastSCOP were adopted to assign a query protein to an evolutionary superfamily.

On average, the fastSCOP took ~3.09 seconds to recognize the structural domain and classification assignment for a single-domain query protein in the query set SCOP-586 (Table 4.1). It was ~338 times faster than MAMMOTH and was ~2.6 times slower than 3D-BLAST, because the fastSCOP required the time of applying MAMMOTH for structure alignments between the query protein and the top ten hit domains. For multiple-domain query proteins, the fastSCOP was ~278 times faster than MAMMOTH and was ~2.7 times slower than 3D-BLAST.

Table 4.1 Accuracy of evolutionary superfamily assignment and average execution time of fastSCOP, 3D-BLAST and MAMMOTH on 586 queries in the set SCOP-586

Query type	Number of queries (Domains)	Program	Number of assigned domains	Assignment accuracy (%)	Unassigned domain percentage (%)	Average time per query (second)	Related to fastSCOP
Single Domain	464 query proteins (464 domains)	3D-BLAST	464	94.4% (95.9%[a])	0%	1.166	0.38
		MAMMOTH	464	98.7% (98.7%[a])	0%	1046.47	338.61
		fastSCOP	455	98.5% (99.6%[a])	1.94%	3.09	1
Multiple Domain	122 query proteins (272 domains)	3D-BLAST	275	86.9%	1.8%	2.238	0.34
		MAMMOTH	238	94.1%	12.5%	1859.80	278.40
		fastSCOP without reassignment[b]	214	98.6%	19.48%	5.11	0.76
		fastSCOP	254	98%	6.6%	6.68	1

[a] Assignment accuracy at SCOP fold level.

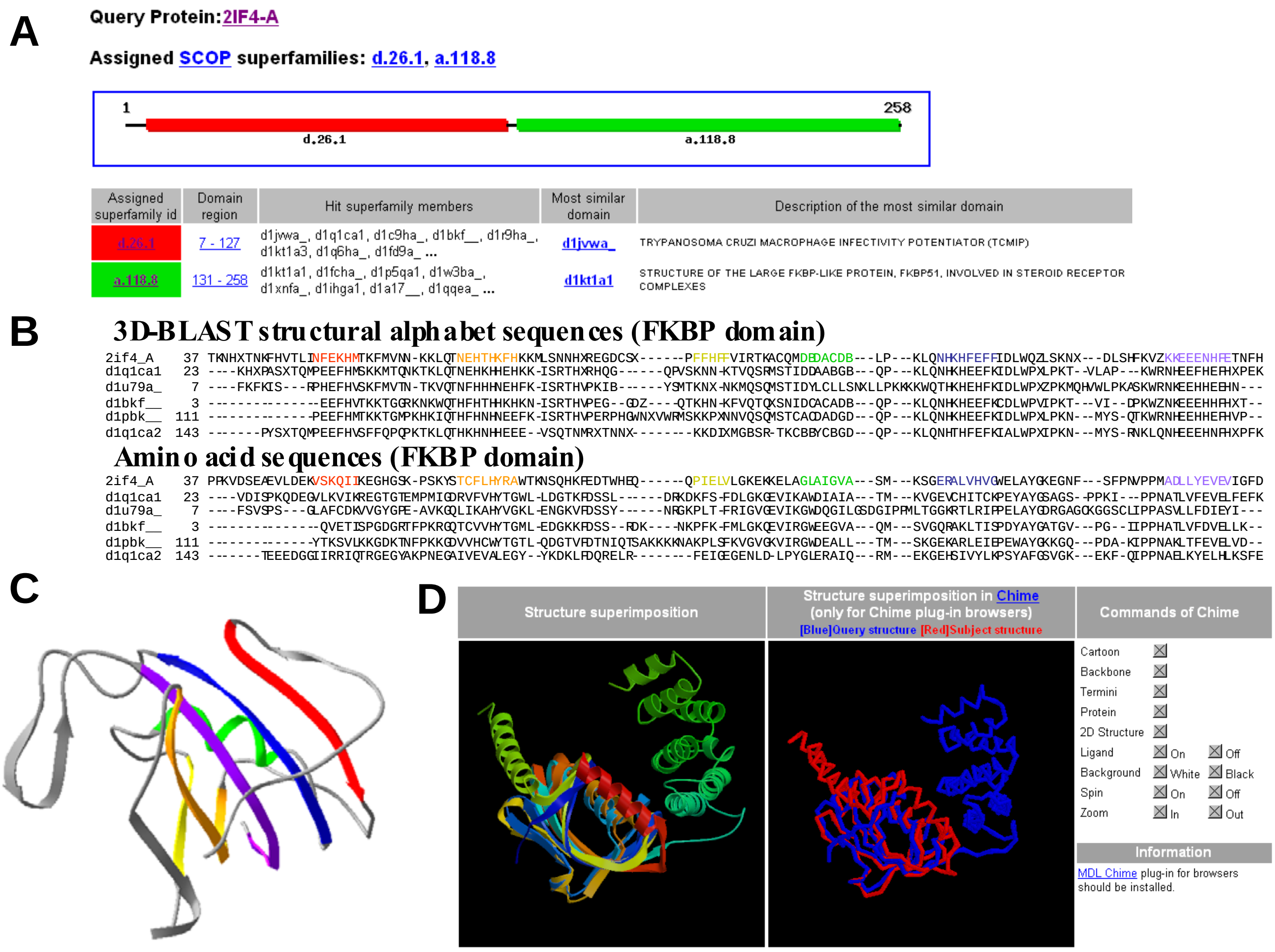
[b] fastSCOP does not apply the reassignment step, which is step 4 in Figure 4.1A.

SCOP-586 consists of 586 query proteins, which are in SCOP1.69 but not in SCOP1.67; the search database is SCOP1.67.

Time was measured using a personal computer with an Intel Pentium 2.8 GHz processor with 1024 Mbytes of RAM.

As shown in Table 4.1, the fastSCOP server yielded 98.5% and 99.6% assignment accuracies at the superfamily and fold levels, respectively, for 464 single-domain queries. It outperformed 3D-BLAST (94.4% and 95.9% at the superfamily and fold levels, respectively) and performed similarly to MAMMOTH (98.7% and 98.7%). The unassignment percentage of the fastSCOP is 1.94% (nine query proteins), which slightly exceeds those of the other two methods. For 122 multiple-domain queries (with 272 domains), the fastSCOP yielded a 98.6% (214 domains) assignment accuracy and the unassignment percentage was 19.48% (53 domains) when the reassignment step (step 4 in Figure 4.1A) was not applied. However, the assignment accuracy was 98% (254 domains) and the unassignment percentage was reduced to 6.6% (18 domains) when the fastSCOP used the reassignment step. The accuracy of fastSCOP significantly exceeded that of MAMMOTH (94.1%) and 3D-BLAST (86.9%); the unassignment percentage was lower than that of MAMMOTH (12.5%, 34 domains).

The fastSCOP was evaluated using the 8700 PDB entries, which have no annotations in the SCOP database, and whose publishing date range from Jan 1, 2006 to Dec 5, 2006. The fastSCOP used these 8700 protein structures as queries, and the search classification database was SCOP 1.71. In this set, 22% (1594 proteins) queries were multi-domain proteins. The fastSCOP server can automatically assign 7311 (84%) proteins (9420 domains) to the SCOP superfamilies in 9.6 hours. According to the assignment accuracy (~98%) of the fastSCOP applied to the query set SCOP-586 and the assignment criteria (step 2 in Figure 4.1A), the fastSCOP server accurately assigns ~9000 domains.



Query Protein: 2IF4-A

Assigned SCOP superfamilies: d.26.1, a.118.8

Assigned superfamily id	Domain region	Hit superfamily members	Most similar domain	Description of the most similar domain
d.26.1	7 - 127	d1jwwa_, d1q1ca1, d1c9ha_, d1bkf__, d1r9ha_, d1kt1a3, d1q6ha_, d1fd9a_ ...	d1jvwa_	TRYPANOSOMA CRUZI MACROPHAGE INFECTIVITY POTENTIATOR (TCMIP)
a.118.8	131 - 258	d1kt1a1, d1fcha_, d1p5qa1, d1w3ba_, d1xnfa_, d1ihga1, d1a17__, d1qqea_ ...	d1kt1a1	STRUCTURE OF THE LARGE FKBP-LIKE PROTEIN, FKBP51, INVOLVED IN STEROID RECEPTOR COMPLEXES

3D-BLAST structural alphabet sequences (FKBP domain)

```
2if4_A    37 TKNHXTNKFHVTLINFEKHMTKFMVNN-KKLQTNEHTHKFHKKMLSNNHXREGDCSX------PFFHFFVIRTKACQMDBDACDB---LP---KLQNHKHFEFFIDLWQZLSKNX---DLSHFKVZKKEEENHFETNFH
d1q1ca1   23 ----KHXPASXTQMPEEFHMSKKMTQNKTKLQTNEHKHHEHKK-ISRTHXRHQG------QPVSKNN-KTVQSRMSTIDDAABGB---QP---KLQNHKHEEFKIDLWPXLPKT--VLAP---KWRNHEEFHEFHXPEK
d1u79a_    7 ----FKFKIS---RPHEFHVSKFMVTN-TKVQTNFHHHNEHFK-ISRTHVPKIB------YSMTKNX-NKMQSQMSTIDYLCLLSNXLLPKKKKWQTHKHEHFKIDLWPXZPKMQHVWLPKASKWRNKEEHHEEHN---
d1bkf__    3 ---------------EEFHVTKKTGGRKNKWQTHFHTHHKHKN-ISRTHVPEG--GDZ----QTKHN-KFVQTQXSNIDCACADB---QP---KLQNHKHEEFKCDLWPVIPKT---VI--DPKWZNKEEEHHFHXT--
d1pbk__  111 --------------PEEFHMTKKTGMPKHKIQTHFHNHNEEFK-ISRTHVPERPHGWNXVWRMSKKPXNNVQSQMSTCAQDADGD---QP---KLQNHKHEEFKIDLWPXLPKN---MYS-QTKWRNHEEHHEFHVP--
d1q1ca2  143 -------PYSXTQMPEEFHVSFFQPQPKTKLQTHKHNHHEEE--VSQTNMRXTNNX-------KKDIXMGBSR-TKCBBYCBGD---QP---KLQNHTHFEFKIALWPXIPKN---MYS-RNKLQNHEEEHNFHXPFK
```

Amino acid sequences (FKBP domain)

```
2if4_A    37 PPKVDSEAEVLDEKVSKQIIKEGHGSK-PSKYSTCFLHYRAWTKNSQHKFEDTWHEQ------QPIELVLGKEKKELAGLAIGVA---SM---KSGERALVHVGWELAYGKEGNF---SFPNVPPMADLLYEVEVIGFD
d1q1ca1   23 ----VDISPKQDEGVLKVIKREGTGTEMPMIGDRVFVHYTGWL-LDGTKFDSSL------DRKDKFS-FDLGKGEVIKAWDIAIA---TM---KVGEVCHITCKPEYAYGSAGS--PPKI---PPNATLVFEVELFEFK
d1u79a_    7 ----FSVSPS---GLAFCDKVVGYGPE-AVKGQLIKAHYVGKL-ENGKVFDSSY------NRGKPLT-FRIGVGEVIKGWDQGILGSDGIPPMLTGGKRTLRIPPELAYGDRGAGCKGGSCLIPPASVLLFDIEYI---
d1bkf__    3 ---------------QVETISPGDGRTFPKRGQTCVVHYTGML-EDGKKFDSS--RDK----NKPFK-FMLGKQEVIRGWEEGVA---QM---SVGQRAKLTISPDYAYGATGV---PG--IIPPHATLVFDVELLK--
d1pbk__  111 --------------YTKSVLKKGDKTNFPKKGDVVHCWYTGTL-QDGTVFDTNIQTSAKKKKNAKPLSFKVGVGKVIRGWDEALL---TM---SKGEKARLEIEPEWAYGKKGQ---PDA-KIPPNAKLTFEVELVD--
d1q1ca2  143 -------TEEEDGGIIRRIQTRGEGYAKPNEGAIVEVALEGY--YKDKLFDQRELR-------FEIGEGENLD-LPYGLERAIQ---RM---EKGEHSIVYLKPSYAFGSVGK---EKF-QIPPNAELKYELHLKSFE
```

Structure superimposition	Structure superimposition in Chime (only for Chime plug-in browsers) [Blue]Query structure [Red]Subject structure	Commands of Chime

Cartoon ☒
Backbone ☒
Termini ☒
Protein ☒
2D Structure ☒
Ligand ☒ On ☒ Off
Background ☒ White ☒ Black
Spin ☒ On ☒ Off
Zoom ☒ In ☒ Out

Information

MDL Chime plug-in for browsers should be installed.

Figure 4.2 Evolutionary superfamily assignment and structural alignment of the fastSCOP server using the structure of multi-domain immunophilin (*At*FKBP42) from *Arabidopsis thaliana* (PDB code 2IF4-A) as the query.

4.3.2 Example analysis

Figure 4.2 shows a fastSCOP result with multi-domain immunophilin (AtFKBP42) from Arabidopsis thaliana (PDB code 2IF4-A) [100] as the query structure. The release date of this protein is Oct 31, 2006, and this protein has not been recorded in SCOP. As shown in Figure 4.2A, the fastSCOP recognized two domains and their SCOP superfamilies, which are the FKBP-like superfamily (SCOP entry d.26.1) and the TPR-like superfamily (SCOP entry a.118.8) for this query. The FKBP domain (Figure 4.2C) of AtFKBP42 consists of a six-stranded anti-parallel β-sheet, wrapped around a short α-helix, and is similar to those of FKBP52 (PDB code 1Q1C-A) [101], FKBP 25 (PDB code 1PBK) [102], FKBP 13 (PDB code 1U79-A) [103] and FKBP 12 (PDB code 1BKF) [104]. The FKBP domain has been demonstrated to interact with plasma membrane-localized ABC transporters AtPGP1 and AtPGP, which directly mediate cellular auxin efflux [105]. The TPR domain of AtFKBP42 is completely helical and binds to AtHSP90, which is critical to plant development and phenotypic plasticity [106, 107].

After the structural domains and evolutionary superfamilies were recognized, the fastSCOP server allowed users to browse similar structures of these superfamilies. Using this AtFKBP42 as a query, the server can identify 13 and 17 similar structures of the FKBP-like domain and TPR domain, respectively. Figure 4.2B illustrates the multiple amino-acid sequence alignment and structural alphabet alignment between AtFKBP42 and five FKBP-like homologous proteins, including FKBP52, FKBP 25, FKBP 13 and FKBP 12. The aligned secondary structures are represented as a continuous color spectrum from red through orange, yellow, green and blue to violet (Figures 4.2B and 4.2C). The structural alphabets were strongly conserved in areas of the secondary structures, which are β-strands (represented by structural alphabets E, F, H, K, and N) or α-helices (represented by structural alphabets A, Y, B, C, and D). These results reveal that the structural alphabet sequences are much better conserved than the amino acid sequences, which result explains why 3D-BLAST detected these distantly related proteins.

4.4 Web service

The fastSCOP server is accessible at “http:// fastSCOP.life.nctu.edu.tw/.” The server can identify the structural domains and determine the evolutionary classification of a query structure from evolutionary classification databases. Users input a PDB code with a protein chain (e.g. 2IF4-A). When the query structure is a new protein structure, the fastSCOP server enables users to input the structure file in PDB format.

This server typically yielded structural domains and the SCOP superfamilies of a query structure in an average of 6 seconds (Figure 4.2A). The server can present the members of the assigned SCOP superfamily and provide both multiple sequence alignments and multiple structural alignments (Figure 4.2B) based on users' requirements. The aligned structures are visualized in PNG format in MolScript and Raster3D packages (Figures 4.2C and 4.2D). The server allows a user to download the aligned structure coordinates in PDB format.

4.5 Summary

This work demonstrated the robustness and feasibility of the fastSCOP server for recognizing the structural domains and the evolutionary classifications of protein structures. The key contribution of this work is the cooperative integration in fastSCOP of 3D-BLAST (a fast structural database search tool) and MAMMOTH (a fast detailed structural alignment tool); the former is required for efficiency and the latter for accuracy. Future works will adopt the fastSCOP for other evolutionary classification databases, such as CATH. Additionally, the fastSCOP can be applied to develop structural motifs and sequence motifs from multiple structure and sequence alignments.

Chapter 5

Conclusions

5.1 Summary

In this thesis, a new approach named 3D-BLAST is proposed for fast structural database searches. The core idea of 3D-BLAST was to design a structural alphabet—to be used to code 3D protein structure databases into structural alphabet sequence databases (SADB)—and a structural alphabet substitution matrix (SASM). We then enhanced the sequence alignment tool BLAST, which searches the SADB using the matrix SASM to rapidly determine protein structure homology or evolutionary classification. 3D-BLAST was designed to maintain the advantages of BLAST, including its robust statistical basis, effective and reliable database search capabilities, and established reputation in biology.

3D-BLAST is rapid and accurate in scanning a large protein structural database, and is useful in an initial scan for similar protein structures, which can be refined using detailed structural comparison methods .However, the use of 3D-BLAST as a search tool also has several limitations, which are (a) 3D-BLAST may have made minor shifts in aligning two local segments with similar letters, (b) the E-values of the hit proteins are insignificant, and (c) the query is a multiple-domain protein. Because of this, an automated server (fastSCOP) is presented, which integrates a fast structure database search tool (3D-BLAST) and a detailed structural alignment tool (MAMMOTH), to recognize SCOP domains and evolutionary superfamilies of a query structure. The classification accuracy of this server is 98% for 464 single-domain queries and 122 multiple-domain queries.

In addition, this study has analyzed the feasibility of studying Space-Related Pharmamotif (SRP) and demonstrated some preliminary results of SRP applied to biosynthesis pathway or cancer pathway. We believe that 3D-BLAST is adopted to develop the motif search tool, called as 3D-PHI-BLAST, for rapidly pharmalogous search.

5.2 Major Contributions

In short, the major contributions of this thesis can be summarized in the following:

1. We have developed a novel kappa-alpha (κ, α) plot derived structural alphabet and a novel BLOSUM-like substitution matrix, called structural alphabet substitution matrix (SASM) which searches in a structural alphabet database (SADB).

2. We present a novel protein structure database search tool, 3D-BLAST, that is useful for analyzing novel structures and can return a ranked list of alignments. This tool has the features of BLAST (for example, robust statistical basis, and effective and reliable search capabilities) and employs a kappa-alpha (κ, α) plot derived structural alphabet and a new substitution matrix. 3D-BLAST searches more than 12,000 protein structures in 1.2 s and yields good results in zones with low sequence similarity.

3. We have built an automated server (fastSCOP), which integrates a fast structure database search tool (3D-BLAST) and a detailed structural comparison tool (MAMMOTH), to recognize SCOP domains and SCOP superfamilies of a query structure. MAMMOTH provided the Z-score and root-mean-square deviation (RMSD) of the C_α atom positions of the aligned residues between the query structure and the hit structure according to the Euclidean distance between corresponding residues rather than the distance between amino acid ‘types’ used in sequence alignments. To combine 3D-BLAST and MAMMOTH is able to reduce the ill effects of 3D-BLAST to improve the assignment accuracy.

5.3 Future Perspectives

5.3.1 Space-Related Pharmamotif discovery in interaction site of protein

Small protein sequence or structural segments with highly conserved properties that may have important biological functions. On the basis of conservation of criteria, like psychochemical property and structural similarity, several conserved segments of proteins belonging to the same protein family with specific function have been identified. These segments are termed ‘structural motifs’. These motifs with their spatial orientation and preservation of structural similarity represent the conserved core of each protein family. Previous studies have been developed for prediction of fold and function of a protein using short segments of sequence and/or structural elements [108-111].

Various methods have been proposed so far for the automated motif discovery in a set of protein sequences [112]. These discovery methods use aligned sequences or multiple sequence alignment (MSA) as an input such as PRINTS [37], PROSITE [38, 113], and Pfam [39]. Besides, TEIRESIAS [40], PRATT2 [41] and a specific pattern growth approach [42] are applied to directly identify frequent patterns from unaligned biological sequences without aligning them. Although motif discovery approaches with unaligned sequence only are more efficiency and less computationally intensive, it may provide the less biological meanings. Subsequently, many of the most functional and evolutionary relationships between homologous protein are so distinct that they cannot be clearly detected through MSA and are evident only by pairwise or multiple structure comparison of the 3D structures. In addition, sequence-based representations are only an approximation to the underlying structural and functional information. Therefore, structural motifs identified at 3D structure level provide significant and reliable information.

A set of functional structural motifs need not to be contiguous in sequence and might discover from the clustering in space of similar side chains coming from different parts of homologous proteins. Finding shared structural motifs in a protein family can be applied to map the interaction site of different proteins with the same partner [114], for locating of the binding site for a common ligand. Besides, sequence and structure motifs have an application in drug design [115] when motifs map to functional sites and ligand binding sites.

In the future, we will propose a novel approach for systems biology and drug design based on the recent developed 3D-BLAST method of protein structural identification [34-36]. We will design new structural motifs that can describe the interacting environment in protein active site named Space-Related Pharmamotif (SRP). The SRP is defined as a set of space-related structural motifs that prefers a set of similar protein sub-site structures consistently interact with ligand, DNA or peptide.

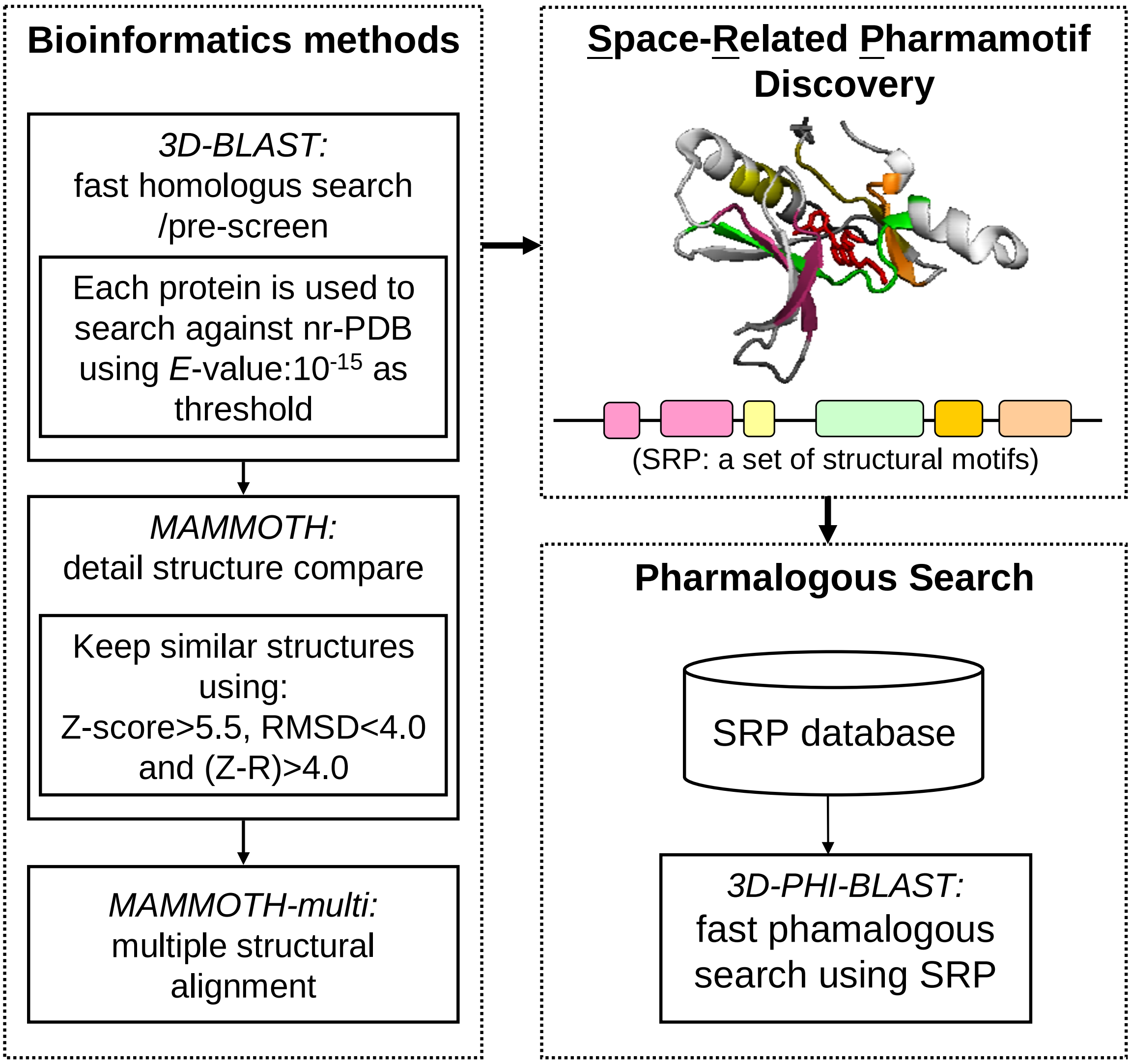

Bioinformatics methods
3D-BLAST:
fast homologus search /pre-screen
Each protein is used to search against nr-PDB using E-value:10-15 as threshold
MAMMOTH:
detail structure compare
Keep similar structures using:
Z-score>5.5, RMSD<4.0 and (Z-R)>4.0
MAMMOTH-multi:
multiple structural alignment
Space-Related Pharmamotif Discovery
(SRP: a set of structural motifs)
Pharmalogous Search
SRP database
3D-PHI-BLAST:
fast phamalogous search using SRP

Figure 5.1 The framework of Space-Related Pharmamotif Discovery and pharmalogs search.

Figure 5.1 shows that the conceptual framework of fast SRP discovery and fast pharmalogs search using SRP. For a group of proteins with similar function and ligand, we build up a set of interacting environment structural motifs and provide fast SRP discovery. Using tertiary protein structure, 3D-BLAST not only allows a fast protein similarity search but also identifies 23 states of the structural alphabet (SA) sequences that represent local structure of SRP. We integrate 3D-BLAST and a detailed structural alignment tool (MAMMOTH [10] and MAMMOTH-multi [116]) to recognize sub-site structures consistently interact with ligand. We use 3D-BLAST to scan quickly the PDB database [4] and selected the homologous structures. MAMMOTH and MAMMOTH-multi was then adopted to align sequentially the query structure with each homologous structure to refine the detailed amino acid position of alignment. Finally, we identify SRP based on the functional or ligand-binding sites of protein and their spatial orientation.

Besides, our novel approach can be applied to fast pharmalogous search using SRP, as named as 3D-PHI-BLAST (Figure 5.1). According to results of the discovery of SRP, we are able to construct SRP with various functions into a database. Using protein with unknown function as query, the 3D-PHI-BLAST may provide rapid motif search through the protein structure and SRP database to predict function and ligand/DNA/peptide pharmacophore binding model.

5.3.2 Immunoinformatics

In the future, 23-state structural alphabet will be aimed to peptide drug design and developing immunoinformatics. For peptide drug design, we will focus in peptide-peptide interaction and build peptide fragment profile database. The peptide fragment profile database will be constructed by 3D-BLAST, our structural motif database and large information about various peptide-peptide interactions.

Besides, we will propose an immunoinformatics system which includes structural immunoinformatics methodology and immunological databases. The system is able to screen and design the antibodies/peptides with high specificity to diagnostic and therapeutic applications. We will develop several structural bioinformatics methods and enhance/modify them for immunology purpose. We will build the integrated immunological databases which include CDR segment database, epitope database and CDR-Epitope interactions database. Additionally, we will offer services for searching between these databases and present the statistical significance of a search to indicate the reliability of the prediction. Furthermore, we will develop an antibody selection platform as the practical application. In this platform, this platform will be combined with phage-display library and yeast cell-display library. Also, the antibody selection platform provides rapid motif search to predict therapeutic peptide and visualization of drug selection.

Bibliography

1. Burley, S.K., et al., *Structural genomics: beyond the human genome project.* Nature Genetics, 1999. **23**: p. 151-157.
2. Burley, S.K. and J.B. Bonanno, *Structural genomics of proteins from conserved biochemical pathways and processes.* Current Opinion in Structural Biology, 2002. **12**: p. 383-391.
3. Todd, A.E., et al., *Progress of structural genomics initiatives: an analysis of solved target structures.* Journal of Molecular Biology, 2005. **348**: p. 1235-1260.
4. Deshpande, N., et al., *The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema.* Nucleic Acids Research, 2005. **33**: p. D233-D237.
5. Watson, J.D., R.A. Laskowski, and J.M. Thornton, *Predicting protein function from sequence and structural data.* Current Opinion in Structural Biology, 2005. **15**: p. 275-284.
6. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **215**: p. 403-410.
7. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Research, 1997. **25**: p. 3389-3402.
8. Holm, L. and C. Sander, *Protein structure comparison by alignment of distance matrices.* Journal of Molecular Biology, 1993. **233**: p. 123-138.
9. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.* Protein Engineering, 1998. **11**: p. 739-747.
10. Ortiz, A.R., C.E. Strauss, and O. Olmea, *MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison.* Protein Science, 2002. **11**: p. 2606-2621.
11. Madej, T., J.F. Gibrat, and S.H. Bryant, *Threading a database of protein cores.* Proteins, 1995. **23**: p. 356-369.
12. Aung, Z. and K.L. Tan, *Rapid 3D protein structure database searching using information retrieval techniques.* Bioinformatics, 2004. **20**: p. 1045-1052.
13. Shyu, C.R., et al., *ProteinDBS: a real-time retrieval system for protein structure comparison.* Nucleic Acids Research, 2004. **32**: p. W572-W575.
14. Martin, A.C., *The ups and downs of protein topology; rapid comparison of protein structure.* Protein Engineering, 2000. **13**: p. 829-837.
15. Guyon, F., et al., *SA-Search: a web tool for protein structure mining based on a Structural Alphabet.* Nucleic Acids Research, 2004. **32**: p. W545-W548.
16. Carpentier, M., S. Brouillet, and J. Pothier, *YAKUSA: a fast structural database*

scanning method. Proteins, 2005. **61**: p. 137-151.

17. Bystroff, C. and D. Baker, *Prediction of local structure in proteins using a library of sequence-structure motifs.* Journal of Molecular Biology, 1998. **281**: p. 565-577.

18. Camproux, A.C., R. Gautier, and P. Tuffery, *A hidden markov model derived structural alphabet for proteins.* Journal of Molecular Biology, 2004. **339**: p. 591-605.

19. de Brevern, A.G., C. Etchebest, and S. Hazout, *Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks.* Proteins, 2000. **41**: p. 271-287.

20. Fetrow, J.S., M.J. Palumbo, and G. Berg, *Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme.* Proteins, 1997. **27**: p. 249-271.

21. Kolodny, R., et al., *Small libraries of protein fragments model native protein structures accurately.* Journal of Molecular Biology, 2002. **323**: p. 297-307.

22. Levitt, M., *Accurate modeling of protein conformation by automatic segment matching.* Journal of Molecular Biology, 1992. **226**: p. 507-533.

23. Rooman, M.J., J. Rodriguez, and S.J. Wodak, *Automatic definition of recurrent local structure motifs in proteins.* Journal of Molecular Biology, 1990. **213**: p. 327-336.

24. de Brevern, A.G., *New assessment of a structural alphabet.* In Silico Biol, 2005. **5**: p. 283-289.

25. Tyagi, M., et al., *A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications.* Proteins, 2006. **65**: p. 32-39.

26. Tyagi, M., et al., *Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet.* Nucleic Acids Res, 2006. **34**: p. W119-W123.

27. Unger, R. and J.L. Sussman, *The importance of short structural motifs in protein structure analysis.* J Comput Aided Mol Des, 1993. **7**: p. 457-472.

28. Fourrier, L., C. Benros, and A.G. de Brevern, *Use of a structural alphabet for analysis of short loops connecting repetitive structures.* BMC Bioinformatics, 2004. **5**: p. 58.

29. Lo, W.C., et al., *Protein structural similarity search by Ramachandran codes.* BMC Bioinformatics, 2007. **8**: p. 307.

30. Lo, W.C., et al., *iSARST: an integrated SARST web server for rapid protein structural similarity searches.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W545-51.

31. Ramachandran, G.N. and V. Sasisekharan, *Conformation of polypeptides and proteins.* Adv Protein Chem, 1968. **23**: p. 283-438.

32. Lo, W.C. and P.C. Lyu, *CPSARST: an efficient circular permutation search tool applied to the detection of novel protein structural relationships.* Genome Biol, 2008. **9**(1): p. R11.

33. Chotia, C. and A.M. Lesk, *The relation between the divergence of sequence and structure in proteins.* EMBO J., 1986. **5**: p. 823-826.

34. Tung, C.H., J.W. Huang, and J.M. Yang, *Kappa-alpha plot derived structural alphabet*

and BLOSUM-like substitution matrix for rapid search of protein structure database. Genome Biology, 2007. **8**: p. R31.1-R31.16.

35. Yang, J.M. and C.H. Tung, *Protein structure database search and evolutionary classification.* Nucleic Acids Research, 2006. **34**: p. 3646-3659.

36. Tung, C.H. and J.M. Yang, *fastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies.* Nucleic Acids Research, 2007. **35**: p. W438-W443.

37. Attwood, T.K., et al., *PRINTS and its automatic supplement, prePRINTS.* Nucleic Acids Research, 2003. **31**: p. 400-402.

38. Hulo, N., et al., *The PROSITE database.* Nucleic Acids Research, 2006. **34**: p. D227-D230.

39. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Research, 2004. **32**: p. D138-D141.

40. Rigoutsos, I. and A. Floratos, *Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm.* Bioinformatics, 1998. **14**: p. 55-67.

41. Jonassen, I., J.F. Collins, and D.G. Higgins, *Finding flexible patterns in unaligned protein sequences.* Protein Science, 1995. **4**: p. 1587-1595.

42. Ye, K., W.A. Kosters, and A.P. Ijzerman, *An efficient, versatile and scalable pattern growth approach to mine frequent patterns in unaligned protein sequences.* Bioinformatics, 2007. **23**: p. 687-693.

43. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* Journal of Molecular Biology, 1995. **247**: p. 536-540.

44. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proceedings of the National Academy of Sciences of the United States of America, 1992. **89**: p. 10915-10919.

45. Huang, C.C., et al., *Structural basis of tyrosine sulfation and VH-gene usage in antibodies that recognize the HIV type 1 coreceptor-binding site on gp120.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**: p. 2706-2711.

46. Adachi, S., et al., *Direct observation of photolysis-induced tertiary structural changes in hemoglobin.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**: p. 7039-7044.

47. Takano, K., Y. Yamagata, and K. Yutani, *Role of amino acid residues at turns in the conformational stability and folding of human lysozyme.* Biochemistry, 2000. **39**: p. 8655-8665.

48. Hutchinson, E.G. and J.M. Thornton, *PROMOTIF--a program to identify and analyze structural motifs in proteins.* Protein Science, 1996. **5**: p. 212-220.

49. Banner, D.W., et al., *Atomic coordinates for triose phosphate isomerase from chicken*

muscle. Biochemical and Biophysical Research Communications, 1976. **72**: p. 146-155.

50. Hogbom, M., et al., *The radical site in chlamydial ribonucleotide reductase defines a new R2 subclass.* Science, 2004. **305**: p. 245-248.

51. Kumar, S. and M. Bansal, *Geometrical and sequence characteristics of alpha-helices in globular proteins.* Biophysical Journal, 1998. **75**: p. 1935-1944.

52. Barlow, D.J. and J.M. Thornton, *Helix geometry in proteins.* Journal of Molecular Biology, 1988. **201**: p. 601-619.

53. Milner-White, E.J., *Recurring loop motif in proteins that occurs in right-handed and left-handed forms. Its relationship with alpha-helices and beta-bulge loops.* Journal of Molecular Biology, 1988. **199**: p. 503-511.

54. Pearson, W.R. and D.J. Lipman, *Improved tools for biological sequence comparison.* Proceedings of the National Academy of Sciences of the United States of America, 1988. **85**: p. 2444-2448.

55. Karplus, K., C. Barrett, and R. Hughey, *Hidden Markov models for detecting remote protein homologies.* Bioinformatics, 1998. **14**: p. 846-856.

56. Pearl, F., et al., *The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis.* Nucleic Acids Research, 2005. **33**: p. D247-D251.

57. Vetting, M.W., et al., *A bacterial acetyltransferase capable of regioselective N-acetylation of antibiotics and histones.* Chemistry & Biology, 2004. **11**: p. 565-573.

58. Nagano, N., C.A. Orengo, and J.M. Thornton, *One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions.* Journal of Molecular Biology, 2002. **321**: p. 741-765.

59. Wolf, E., et al., *Crystal structure of a GCN5-related N-acetyltransferase: Serratia marcescens aminoglycoside 3-N-acetyltransferase.* Cell, 1998. **94**: p. 439-449.

60. Peapus, D.H., et al., *Structural characterization of the enzyme-substrate, enzyme-intermediate, and enzyme-product complexes of thiamin phosphate synthase.* Biochemistry, 2001. **40**: p. 10103-10114.

61. Terwilliger, T.C., *Structural genomics in North America.* Nature Structural Biology, 2000. **7 Suppl**: p. 935-939.

62. Wilmanns, M., et al., *Structural conservation in parallel beta/alpha-barrel enzymes that catalyze three sequential reactions in the pathway of tryptophan biosynthesis.* Biochemistry, 1991. **30**: p. 9161-9169.

63. Schaffer, A.A., et al., *IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices.* Bioinformatics, 1999. **15**: p. 1000-1011.

64. Uzumaki, T., et al., *Crystal structure of the C-terminal clock-oscillator domain of the cyanobacterial KaiA protein.* Nature Structural & Molecular Biology, 2004. **11**: p.

623-631.

65. Lamb, A.L., et al., *Heterodimeric structure of superoxide dismutase in complex with its metallochaperone.* Nature Structural Biology, 2001. **8**: p. 751-755.

66. Rosenzweig, A.C., et al., *Crystal structure of the Atx1 metallochaperone protein at 1.02 A resolution.* Structure, 1999. **7**: p. 605-617.

67. Hurley, J.K., et al., *Structure-function relationships in Anabaena ferredoxin: correlations between X-ray crystal structures, reduction potentials, and rate constants of electron transfer to ferredoxin:NADP+ reductase for site-specific ferredoxin mutants.* Biochemistry, 1997. **36**: p. 11100-11117.

68. Zhang, C. and S.H. Kim, *Overview of structural genomics: from structure to function.* Current Opinion in Chemical Biology, 2003. **7**: p. 28-32.

69. Chance, M.R., et al., *High-throughput computational and experimental techniques in structural genomics.* Genome Research, 2004. **14**: p. 2145-2154.

70. Grandori, R. and J. Carey, *Six new candidate members of the alpha/beta twisted open-sheet family detected by sequence similarity to flavodoxin.* Protein Science, 1994. **3**: p. 2185-2193.

71. Frazao, C., et al., *Structure of a dioxygen reduction enzyme from Desulfovibrio gigas.* Nature Structural Biology, 2000. **7**: p. 1041-1045.

72. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Research, 2004. **32**: p. D258-D261.

73. Falquet, L., et al., *The PROSITE database, its status in 2002.* Nucleic Acids Research, 2002. **30**: p. 235-238.

74. Park, C.J., et al., *Solution structure of the influenza A virus cRNA promoter: implications for differential recognition of viral promoter structures by RNA-dependent RNA polymerase.* Nucleic Acids Research, 2003. **31**: p. 2824-2832.

75. Madrona, A.Y. and D.K. Wilson, *The structure of Ski8p, a protein regulating mRNA degradation: Implications for WD protein structure.* Protein Science, 2004. **13**: p. 1557-1565.

76. Heuer, K., et al., *Structure of a helically extended SH3 domain of the T cell adapter protein ADAP.* Structure, 2004. **12**: p. 603-610.

77. Mowat, C.G., et al., *Octaheme tetrathionate reductase is a respiratory enzyme with novel heme ligation.* Nature Structural & Molecular Biology, 2004. **11**: p. 1023-1024.

78. Xu, X.F., et al., *NMR structure of a type IVb pilin from Salmonella typhi and its assembly into pilus.* The Journal of Biological Chemistry, 2004. **279**: p. 31599-31605.

79. Kusunoki, H., R.I. MacDonald, and A. Mondragon, *Structural insights into the stability and flexibility of unusual erythroid spectrin repeats.* Structure, 2004. **12**: p. 645-656.

80. Rubin, S.M., et al., *Solution structure of a putative ribosome binding protein from Mycoplasma pneumoniae and comparison to a distant homolog.* Journal of Structural

and Functional Genomics, 2003. **4**: p. 235-243.

81. Turner, C.F. and P.B. Moore, *The solution structure of ribosomal protein L18 from Bacillus stearothermophilus.* Journal of Molecular Biology, 2004. **335**: p. 679-684.

82. Landrieu, I., et al., *A small CDC25 dual-specificity tyrosine-phosphatase isoform in Arabidopsis thaliana.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**: p. 13380-13385.

83. Hickman, A.B., et al., *The structural basis of ordered substrate binding by serotonin N-acetyltransferase: enzyme complex at 1.8 A resolution with a bisubstrate analog.* Cell, 1999. **97**: p. 361-369.

84. Roll-Mecak, A., et al., *X-Ray structures of the universal translation initiation factor IF2/eIF5B: conformational changes on GDP and GTP binding.* Cell, 2000. **103**: p. 781-792.

85. Liu, S., et al., *Structures of human dihydroorotate dehydrogenase in complex with antiproliferative agents.* Structure, 2000. **8**: p. 25-33.

86. Hall, L.T., et al., *X-ray crystallographic and analytical ultracentrifugation analyses of truncated and full-length yeast copper chaperones for SOD (LYS7): a dimer-dimer model of LYS7-SOD association and copper delivery.* Biochemistry, 2000. **39**: p. 3611-3623.

87. Hu, Y., et al., *Crystal structure of S-adenosylhomocysteine hydrolase from rat liver.* Biochemistry, 1999. **38**: p. 8323-8333.

88. Lindahl, E. and A. Elofsson, *Identification of related proteins on family, superfamily and fold level.* Journal of Molecular Biology, 2000. **295**: p. 613-625.

89. Eddy, S.R., *Profile hidden Markov models.* Bioinformatics, 1998. **14**: p. 755-763.

90. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Research, 2003. **31**: p. 365-370.

91. Schneider, R. and C. Sander, *The HSSP database of protein structure-sequence alignments.* Nucleic Acids Research, 1996. **24**: p. 201-205.

92. Buglino, J., et al., *Crystal structure of PapA5, a phthiocerol dimycocerosyl transferase from Mycobacterium tuberculosis.* The Journal of Biological Chemistry, 2004. **279**: p. 30634-30642.

93. Singer, A.U., et al., *Crystal structures of the type III effector protein AvrPphF and its chaperone reveal residues required for plant pathogenesis.* Structure, 2004. **12**: p. 1669-1681.

94. Weljie, A.M. and H.J. Vogel, *Unexpected structure of the Ca2+-regulatory region from soybean calcium-dependent protein kinase-alpha.* The Journal of Biological Chemistry, 2004. **279**: p. 35494-35502.

95. Rak, A., et al., *Structure of Rab GDP-dissociation inhibitor in complex with prenylated YPT1 GTPase.* Science, 2003. **302**: p. 646-650.

96. Andreeva, A., et al., *SCOP database in 2004: refinements integrate structure and*

sequence family data. Nucleic Acids Research, 2004. **32**: p. D226-D229.

97. Cheek, S., et al., *SCOPmap: automated assignment of protein structures to evolutionary superfamilies.* BMC Bioinformatics, 2004. **5**: p. 197.

98. Gough, J., et al., *Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.* Journal of Molecular Biology, 2001. **313**: p. 903-919.

99. Krissinel, E. and K. Henrick, *Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions.* Acta Crystallographica. Section D, Biological Crystallography, 2004. **60**: p. 2256-2268.

100. Granzin, J., A. Eckhoff, and O.H. Weiergraber, *Crystal structure of a multi-domain immunophilin from Arabidopsis thaliana: a paradigm for regulation of plant ABC transporters.* Journal of Molecular Biology, 2006. **364**: p. 799-809.

101. Wu, B., et al., *3D structure of human FK506-binding protein 52: implications for the assembly of the glucocorticoid receptor/Hsp90/immunophilin heterocomplex.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**: p. 8348-8353.

102. Liang, J., et al., *Structure of the human 25 kDa FK506 binding protein complexed with rapamycin.* Journal of the American Chemical Society, 1996. **118**: p. 1231-1232.

103. Gopalan, G., et al., *Structural analysis uncovers a role for redox in regulating FKBP13, an immunophilin of the chloroplast thylakoid lumen.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**: p. 13945-13950.

104. Itoh, S., et al., *Conformation of Fk506 in X-ray structures of its complexes with human recombinant Fkbp12 mutants.* Bioorganic & Medicinal Chemistry Letters, 1995. **5**: p. 1983-1988.

105. Geisler, M., et al., *TWISTED DWARF1, a unique plasma membrane-anchored immunophilin-like protein, interacts with Arabidopsis multidrug resistance-like transporters AtPGP1 and AtPGP19.* Molecular Biology of the Cell, 2003. **14**: p. 4238-4249.

106. Sangster, T.A. and C. Queitsch, *The HSP90 chaperone complex, an emerging force in plant development and phenotypic plasticity.* Current Opinion in Plant Biology, 2005. **8**: p. 86-92.

107. Scheufler, C., et al., *Structure of TPR domain-peptide complexes: critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine.* Cell, 2000. **101**: p. 199-210.

108. Shepherd, A.J., D. Gorse, and J.M. Thornton, *Prediction of the location and type of beta-turns in proteins using neural networks.* Protein Science, 1999. **8**: p. 1045-1055.

109. Kaur, H. and G.P. Raghava, *A neural network method for prediction of beta-turn types in proteins using evolutionary information.* Bioinformatics, 2004. **20**: p. 2751-2758.

110. Schwabe, C., *The structure and evolution of alpha/beta barrel proteins.* The FASEB

Journal, 1996. **10**: p. 184.

111. Kannan, N., et al., *Clusters in alpha/beta barrel proteins: implications for protein structure, function, and folding: a graph theoretical approach.* Proteins, 2001. **43**: p. 103-112.

112. Sandve, G.K. and F. Drablos, *A survey of motif discovery methods in an integrated framework.* Biology Direct, 2006. **1**: p. 11.

113. Sigrist, C.J., et al., *ProRule: a new database containing functional and structural information on PROSITE profiles.* Bioinformatics, 2005. **21**: p. 4060-4066.

114. Marcatili, P., G. Bussotti, and A. Tramontano, *The MoVIN server for the analysis of protein interaction networks.* BMC Bioinformatics, 2008. **9 Suppl 2**: p. S11.

115. Craik, D.J., N.L. Daly, and C. Waine, *The cystine knot motif in toxins and implications for drug design.* Toxicon, 2001. **39**: p. 43-60.

116. Lupyan, D., A. Leo-Macias, and A.R. Ortiz, *A new progressive-iterative algorithm for multiple structure alignment.* Bioinformatics, 2005. **21**: p. 3255-3263.

Appendix A
Standalone 3D-BLAST program

The package can be downloaded from http://3d-blast.life.nctu.edu.tw/download.php.
Also, you may download the package from Standalone_3d-blast_Linux_beta102.tar.gz.
After downloading the package to you Linux-based computer, uncompress it by following commands in terminal.

1. gunzip Standalone_3d-blast_Linux_beta102.tar.gz
2. tar -xpf Standalone_3d-blast_Linux_beta102.tar

And then, you may check the file "README" in the directory "Standalone_3d-blast_Linux" for more information about compilation and usage of 3D-BLAST.

INSTALLATION

============

Contents of the package

----

1. 3d-blast.c	-	The source code of 3D-BLAST
2. path.h	-	The path configuration file
3. Makefile	-	The compilation file
4. data/BLOSUM62	-	Structural alphabet substitution matrix
5. blast/bin/blastp	-	NCBI-BLAST binary
6. blast/bin/formatdb	-	NCBI-FORMATDB binary
7. dsspcmbi/dsspcmbi	-	CMBI-DSSP binary
8. example/SCOP_173_40	-	The example of Structural alphabet database
10. example/example1.pdb	-	The example of protein file in PDB format
11. example/example2.dssp	-	The example of protein file in DSSP format
12. example/SADB_list	-	The example of list file for generating database
13. README	-	This document

Compilation

User make the program with:

```
make -f Makefile
```

This produces the executable file 3D-BLAST.

USAGE

=============

1. Formatting Structural Alphabet DataBase

Before using 3d-blast, user needs to download the structural alphabet database (SADB) in FASTA format from the following link, and format the database using the program "formatdb" from NCBI.

http://3d-blast.life.nctu.edu.tw/download.php

The following command line formats the SADB. The results are saved in various files, including phr, pin, psd, psi, and psq.

```
./3d-blast -db <SADB file>
```

where "<SADB file>" is the path and name of SADB file.

For example,

```
./3d-blast -db example/SCOP_173_40
```

2. Running 3D-BLAST to search structural database

This program searches a protein query with pdb or dssp format against a protein database. If a pdb file is as a query, it first transform the pdb-style file into dssp-style one by using the program "dsspcmbi" from CMBI. And then, it translates the protein 3D structure in 1D Structural Alphabet (SA) sequence. The primary use of 3D-BLAST search is to identify the SA sequence by finding if match(es) are present in the SADB.

In the example command line below, 3D-BLAST searchs the <query protein file> with <chain id> against <SADB file>. The result is saved in <output file>.

```
./3d-blast -p <query protein file> <chain id> -d <SADB file> -o <output file>
```

There are two examples to demostrate how use pdb and dssp file as query to search against SADB.

```
./3d-blast -p example/example1.pdb A -d example/SCOP_173_40 -o 3d-blast_output
./3d-blast -p example/example2.dssp A -d example/SCOP_173_40 -o 3d-blast_output2
```

Optional arguments

-i <Temporary SA sequence file> [String] (default = Temp_SA.seq)

-v <Number of sequences to show one-line descriptions> [Integer] (default = 50)

-b <Number of sequences to show alignments> [Integer] (default = 50)

-e <E-value threshold> [Real] (default = 10.0)

For instance,

```
./3d-blast -p example/example1.pdb A -d example/SCOP_173_40 -o 3d-blast_output -i Other_SA.seq
./3d-blast -p example/example1.pdb A -d example/SCOP_173_40 -o 3d-blast_output -v 10 -b 10
./3d-blast -p example/example1.pdb A -d example/SCOP_173_40 -o 3d-blast_output -e 1e-10
```

3. Generating Structural Alphabet DataBase

There is another way to produce user's SADB instead of downloading it from 3d-blast website. The following command means the 3d-blast program reads a list of pdb-style or dssp-style files with chain id, and then translates all of them into the output of SADB file. The format of the list file in each line is just like "<query protein file> <chain id>" including the names of pdb/dssp file with the path to a directory and the chain id.

```
./3d-blast -mkdb <list file> -o <output SADB file>
```

For example,

```
./3d-blast -mkdb example/SADB_list -o Other_SADB
```

After generating the SADB, user still have to format the SADB as the description of step 1.

4. Generating Structural Alphabet sequence only

This program also provides the function of translating protein structure into SA sequence by using the following command lines. It is also useful to build the custom SADB.

```
./3d-blast -sq_write <query protein> <chain id> -o <output file>
./3d-blast -sq_append <query protein> <chain id> -o <output file>
```

Note that the first command line is to write SA sequence in customSADB and second line is append SA sequence to the same SADB file. For example,

```
./3d-blast -sq_write example/example1.pdb A -o customSADB
./3d-blast -sq_append example/example2.dssp A -o customSADB
```

5. Printing HELP message

It shows the usage message of Standalone 3D-BLAST by following command.

```
./3d-blast -h
./3d-blast -?
```