3646–3659 *Nucleic Acids Research, 2006, Vol. 34, No. 13*
doi:10.1093/nar/gkl395

Protein structure database search and evolutionary classification

Jinn-Moon Yang[1,2,3,*] and Chi-Hua Tung[2]

[1]Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, 30050, Taiwan, [2]Institute of Bioinformatics, National Chiao Tung University, Hsinchu, 30050, Taiwan and [3]Core Facility for Structural Bioinformatics, National Chiao Tung University, Hsinchu, 30050, Taiwan

Received March 8, 2006; Revised May 6, 2006; Accepted May 9, 2006

ABSTRACT

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 (SASM) 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 s and achieved a good agreement with search results from detailed structure alignment methods. [3D-BLAST is available at http://3d-blast.life.nctu.edu.tw]

INTRODUCTION

Genome sequencing projects are in progress for more than 644 organisms, and complete sequences are now available for more than 160 prokaryotic and eukaryotic genomes. In these sequenced genomes, a large portion (~30 to 50%) of genes encode proteins of unknown biological functions. To address this issue, structural genomics is emerging as a powerful approach to assign functional annotations by determining the conformations of numerous proteins in a genome-wide strategy (1–3). Structural genomics projects are generating new structures at an unprecedented rate—a benefit of recent developments in high-throughput technologies. As a result, the number of proteins with unassigned functions and the number of protein structures in the Protein Data Bank (PDB) are increasing rapidly (4).

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), 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 (7), CE (8), MAMMOTH (9) and VAST (10), 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 (11) and ProteinDBS (12) 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 (13), SA-Search (14) and YAKUSA (15), describe protein structures as 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.

We have created a fast protein structure search tool, 3D-BLAST, that can search >10 000 structures in 1.3 s

*To whom correspondence should be addressed. Tel: 886 3 5712121, ext. 56942; Fax: 886 3 5729288; Email: moon@faculty.nctu.edu.tw

using only a personal computer. This innovative program dispenses with the need to perform searches for Euclidean distances between corresponding residues; instead, the highly regarded local sequence alignment tool, BLAST, is used to discover homologous proteins and to evaluate the statistical significance of hits by providing *E*-values from structure databases. Our method encodes 3D protein structures into structural alphabet sequences by mapping 5mer structural segments into corresponding structural letters (14–18). These structural alphabet sequences and our new structural alphabet substitution matrix (SASM) enhance the ability of BLAST to search structural homology of a query sequence to a known protein or family of proteins, often providing clues about the function of a query protein. The 3D-BLAST method is illustrated in Figure 1.

3D-BLAST has the advantages of BLAST for fast structural database scanning and evolutionary classification. It searches for the longest common substructures, called structural alphabet high-scoring segment pairs (SAHSPs), existing between the query structure and every structure in the structural database. The SAHSP is similar to the high-scoring segment pair (HSP) of BLAST (6), which is used to search amino acid sequences. 3D-BLAST ranks the search homology structures based on both SAHSP and *E*-values, which are calculated from the SASM. 3D-BLAST is much faster than related programs and it is available at http://3d-blast.life.nctu.edu.tw.

MATERIALS AND METHODS

Figure 1 presents details of our approach 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 SADB—and a 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. To develop the structural alphabet and the SASM matrix, we prepared 674 structural pairs, each of which had high structural similarity but low sequence identity.

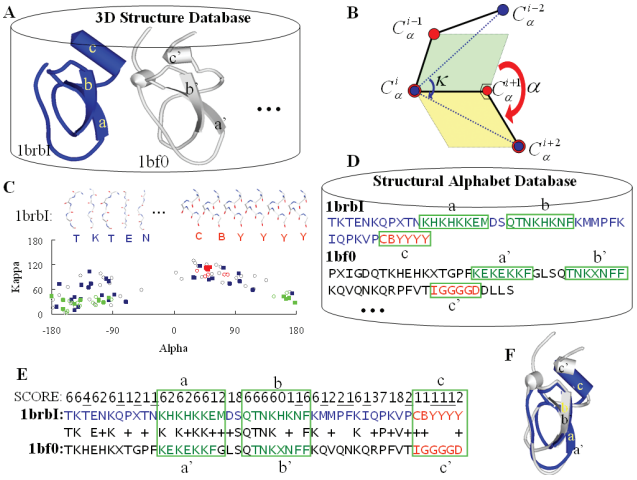


Figure 1. Stepwise illustration of 3D-BLAST using the protein 1brb chain I as the query protein. (**A**) A known 3D database with two structures, 1brbI (blue) and 1bf0 (gray). (**B**) Definition of the κ and α angles. Angle κ, ranging from 0 to 180°, of residue *i* is a bond angle formed by three C_α atoms of residues $i-2$, i and $i+2$. Angle α, 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 $i+2$. (**C**) The (κ, α) maps of 1brbI (square) and 1bf0 (circle) are similar. The strands (green) and helices (red) are indicated. The 3D structure fragments of the first five and last six fragments of 1brbI are given. (**D**) The SADB. (**E**) The result of aligning these two structural alphabet sequences using BLAST and the score of the SASM. For example, the score of aligning T to T is 6, K to K is 6, and T to H is −4. (**F**) The resulting structure alignment.

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. However, the use of BLAST as a search tool also has several limitations, which are the maximum state (19) of the structural alphabet, the need for a new SASM, and a new *E*-value threshold to indicate the statistical significance of an alignment. Furthermore, 3D-BLAST is slow if the structural alphabet is un-normalized, because the BLAST algorithm searches a statistically significant alignment by two main steps (6). It first scans the database for hit words that score more than a threshold value if aligned with words in the query sequence; it then extends each hit word in both directions to check the alignment score. To reduce the ill effect of 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 γ value was set to 16 000 (∼7.0% of total structural segments in the pair database).

Pair database

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 (20). 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 root-mean-square deviation (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.

(κ, α) map

A structure fragment (five residues long) was defined by the (κ, α)-pair angles as shown in Figure 1B. 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

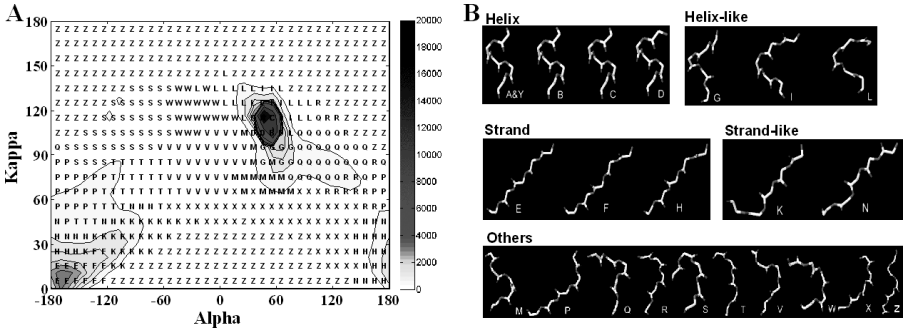


Figure 2. Evaluation of the structural alphabet. (**A**) The accumulated (κ, α) matrix of 225 523 segments derived from the pair database. This matrix, 648 cells (36 × 18), is clustered into 23 groups based on the similarity of representative segments of each cell using a NNC method. A representative structural letter is assigned for each group. (**B**) The representative 3D fragments of 23 structural letters. These 23 segments are roughly divided into five categories: 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.

(κ, α)-map distribution as shown in Figure 1C. The accumulated (κ, α)-map matrix (Figure 2A) 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.

(κ, α)-map cluster and 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 2A). A nearest-neighbor clustering (NNC) algorithm was developed to cluster 225 523 fragments in the accumulated (κ, α)-map matrix (Figure 2A) 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 $w_i = (1/S_i)/(\sum_{j=1}^{M} 1/S_j)$, 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 2B); (5) obtaining a composition (Supplementary Figure S1A) of 23 structural letters i.e. 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).

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 $100 \leqslant \kappa < 105°$ and $40 \leqslant \alpha < 45°$, and the other is in $105 \leqslant \kappa < 110°$ and $45 \leqslant \alpha < 50°$. 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 is given as

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

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}, $1 \leqslant i, \leqslant 647$) 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, $1 \leqslant i \leqslant 20$) 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 20 or all of the cells are labeled. (4) Assign all of the remaining unlabeled cells to a cluster C_{21}.

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 $\sum_{m=1}^{M} w_i w_m D_{im}$, 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

	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 3. SASM of 3D-BLAST. The scores are high if similar letters are aligned (yellow blocks). In contrast, the scores are low when a helix letter is aligned with a strand letter (orange blocks).

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 2A shows the distribution of these 23 clusters and the structural alphabet on 648 cells in the (κ, α)-map. Figure 2B 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.

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 (21). Here, we developed a new SASM matrix (Figure 3) 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 deigned 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 ($1 \leqslant i, j \leqslant 23$), of the SASM matrix is defined as $S_{ij} = \lambda \log_2(q_{ij}/e_{ij})$, 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 $f_{ij}/\sum_{m=1}^{23}\sum_{k=1}^{m} f_{mk}$, 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. In 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; λ is set to 1.89 for the best performance and efficiency. The final score S_{ij} is rounded to the nearest integer value.

RESULTS

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. It returns a list of protein structures that are similar to the query, ordered by *E*-values. When we searched databases, such as SCOP (22) or CATH (23), which are based on structural classification schemes, the evolutionary classification (i.e. 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.

Figure 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 1B) as defined

in the DSSP program (24). Second, as governed by these angles, each structure in the protein structure database has a specific (κ, α)-map distribution (Figure 1C), which was then encoded into a corresponding 1D structural alphabet sequence and stored in the SADB database (Figure 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 1E) is reported along with an *E*-value similar to the one assigned by BLAST, and the structure alignment (Figure 1F) is also reported. For example, the (κ, α)-map distributions (Figure 1C) of 1brbI (filled squares) and 1bf0 (open circles) are similar, as are their protein structures (Figure 1F). In Figure 1C–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%.

Evaluation of the 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 1B, to divide a 3D protein structure into a series of 3D protein fragments.

Figure 2A 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; i.e. the rmsd is <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. This (κ, α) map is similar in concept to the Ramachandran plot, which represents a residue using angles ϕ and ψ.

Based on the characteristics of the (κ, α)-map matrix, we developed a NNC algorithm to cluster these 225 523 3D protein fragments into 23 groups, which are represented by respective structural letters. We found that the structural alphabet can represent the profiles of most of the 3D fragments and be roughly divided into five categories (Figure 2B): 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. For example, the shapes of representative 3D segments in the helix letters are similar and the ones in the strand letters are also similar. The composition of the 23 structural letters (Supplementary Figure S1A) are similar to the one of 20 amino acids in protein sequences. Most of α-helix secondary structures are encoded as the helix or helix-like letters, and none are encoded as the strand or strand-like letters (Supplementary Figure S1B). Moreover, most of β-strand structures are encoded as the strand or strand-like letters.

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 eight letters (the helix and helix-like) for the α-helix structure and five letters (strand and strand-like) for the β-strand structure (Figure 2B). 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*. (18) 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 to 4.5 Å for these 10 proteins and were lightly worse than Kolodny *et al*. (18) work. In the future, we will enhance the structural alphabet for protein structure prediction.

Evaluation of SASM

Substitution matrices are the key component of protein alignment methods. We developed a new SASM (Figure 3) using a method similar to that used to construct BLOSUM62 (21) 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 2B), which allows the peptide backbone to fold back and therefore has

great significance in protein structure and function (19). This substitution score is 11 (Figure 3). Based on the tool PROMOTIF (25), 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 2B) 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 (21). 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. 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.

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. The first 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 second 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 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.

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 $(\sum_{i=1}^{A} i/T_h^i)/A$, where T_h^i is the number of compounds in a hit list containing i correct structures.

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 HSP in BLAST for protein sequence alignment. Therefore, the statistics of HSPs for analyzing the BLAST algorithm allows 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).

Figure 4 and Table 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 4A) and the rate of correct classification increases for evolutionary classification assignment (Figure 4B). For structural database searches, the precision is 0.81 and recall is 0.5 if the *E*-value is $<e^{-15}$ (Table 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 4B). If the *E*-value is $<e^{-15}$, 98.53% of 894 protein structures are assigned correct classifications and the coverage is 91.61% (Table 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 $<e^{-15}$.

Figure 4C shows that 3D-BLAST *E*-values correlate strongly 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

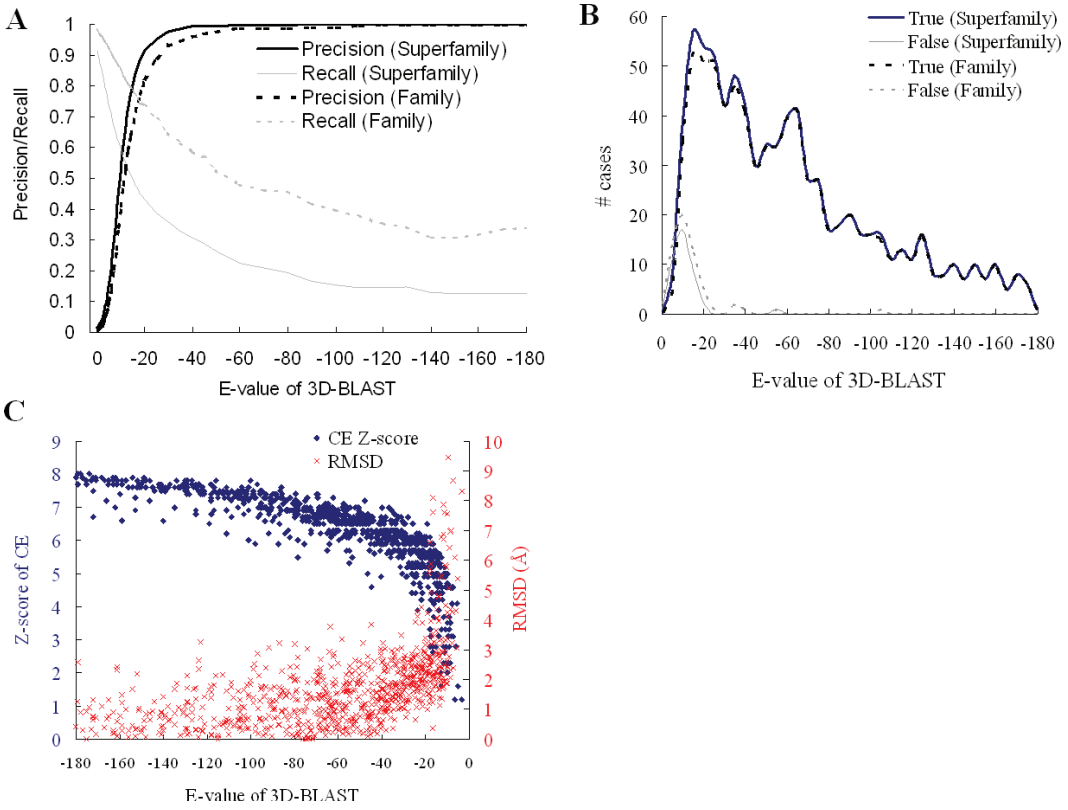


Figure 4. 3D-BLAST performance with *E*-values on the protein query set SCOP-894. (**A**) The relationship between precision and recall for structure database search. The precision is 0.81 and recall is 0.50 if the *E*-value is set to e^{-15}. (**B**) The number of correct and false family/superfamily assignments. The correct percentage of the superfamily assignments is 97.53% if the *E*-value is set to e^{-15}, and the coverage is 91.61%. (**C**) The relationship between 3D-BLAST *E*-values and both Z-Scores of CE (blue) and rmsd of aligned residues (red).

Table 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

Threshold of *E*-value	Structural database search			Superfamily assignment[a]			
	Recall	Precision	False positive rate	894 proteins Correct assignment (%)	Coverage[c](%)	Sequence identity <25%[b] 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, and the coverage is 91.61% if the *E*-value is set to e^{-15} for the query set SCOP-894.

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.

Comparison with PSI-BLAST

Table 2 show 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 (6) 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

Table 2. 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%		
	Number of queries	3D-BLAST Corrected assignment %	PSI-BLAST Corrected assignment %	Number of queries	3D-BLAST Corrected assignment %	PSI-BLAST Corrected assignment %
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 1.

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.

For most sets of sequence identities, 3D-BLAST outperforms PSI-BLAST (Table 2). 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.

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 and 4 show these two cases. The first target is aminoglycoside N-acetyltransferase (NAT) AAC(6′)-Iy (26) (PDB code 1s3z) (Figure 5). The secondary target is a structural genomics target (PDB code 1 × i3) that is a member of a triosephosphate isomerase (TIM) beta/alpha-barrel fold (27) (Figure 6). In each case, 3D-BLAST reported a structurally and functionally relevant relationship in greater detail.

NAT

The *Salmonella enteritidis* NAT AAC(6′)-Iy (PDB code 1s3z) is a member of the GCN5-related *N*-acetyltransferase (GNAT) superfamily (28) 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 5B) (28).

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). 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).

Figure 5A 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 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 5B) 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 5A and two proteins (PDB code 1uth and 1vhs) selected from the NAT family. The

Table 3. 3D-BLAST search results using aminoglycoside 6′-NAT as the query

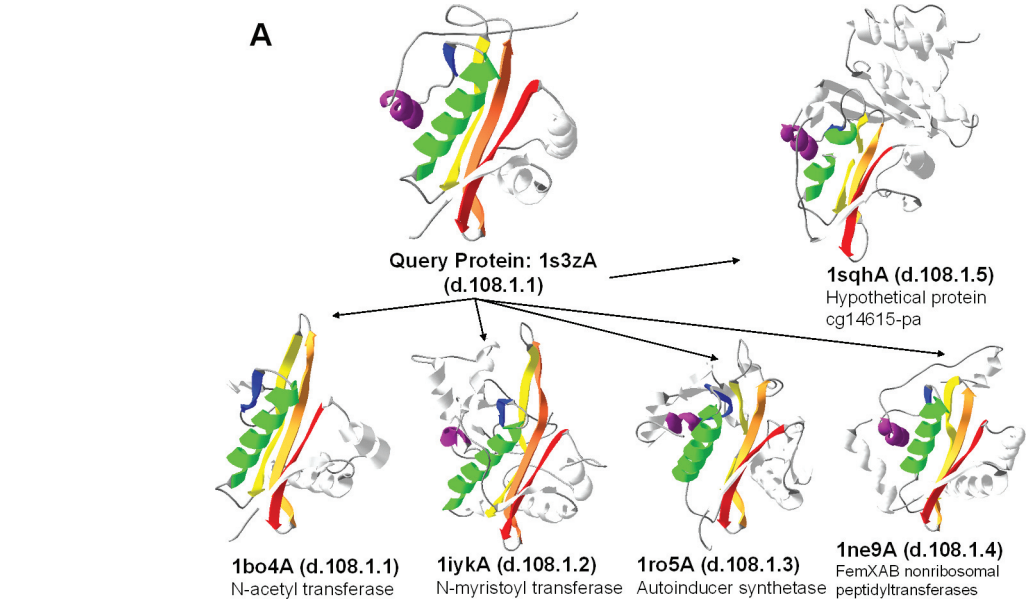
PDB code	Protein name	SCOP family name	log(*E*-value)	Rmsd (Å)	Sequence identity[a]	Species
1tiqA	Protease synthase and sporulation negative regulatory protein PaiA	N-acetyl transferase	−36.70	1.97	17	*Bacillus subtilis*
1qstA	GCN5 histone acetyltransferase	N-acetyl transferase	−32.70	3	14.4	*Tetrahymena thermophila*
1i12A	Glucosamine-phosphate NAT 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	*S.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	*B.subtilis*
1vhsA	Putative phosphinothricin acetyltransferase YwnH	N-acetyl transferase	−26.40	2.68	18.3	*B.subtilis*
1n71A	Aminoglycoside 6′-NAT	N-acetyl transferase	−26.40	2.28	18.8	*Enterococcus faecium*
1m44A	Aminoglycoside 2′-NAT	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	*B.subtilis*
1p0hA[b]	Mycothiol synthase MshD	N-acetyl transferase	−24.30	1.51	14.2	*M.tuberculosis*
1cjwA	Serotonin N-acetyltranferase	N-acetyl transferase	−24.22	3.04	16.6	*Ovis aries*
1bo4A[c]	Aminoglycoside 3-NAT	N-acetyl transferase	**−24.22**	**2.74**	**16.8**	*Serratia marcescens*
1nslA	Probable acetyltransferase YdaF	N-acetyl transferase	−23.52	2.92	18.1	*B.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	*S.cerevisiae*
1q2yA	Probable acetyltransferase YjcF	N-acetyl transferase	−19.70	2.48	19	*B.subtilis*
1bob	Histone acetyltransferase HAT1	N-acetyl transferase	−16.15	2.18	14.9	*S.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	*S.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	*S.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	*S.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	*C.albicans*
1lrzA2	Methicillin resistance protein FemA	FemXAB	−11.52	3.46	16.7	*S.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 5A.

Table 4. Structure database search results of 3D-BLAST for finding homologous superfamilies in SCOP 95 using thiamine phosphate pyrophosphorylase from *P.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
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.



B

3D-BLAST structural alphabet sequences

```
                    10        20        30        40        50        60        70        80        90        100       110       120
                     |         |         |         |         |         |         |         |         |         |         |         |
SS structure SSSSSS       SSSSSSS  SSSS                    SSS  SSSSSSS          HHHHHHHHHHHHH          SSSSS       HHHHHHHH
Motif        DDDDDDDDDDDDDDDDDDDDDDDDD                     AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA       BBBBBBBBBBBBBBBBBBBBBBBBBBBB
             +++++     +  +++++++                           +  +++++++   ++++   +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 5. Multiple structure alignment and multiple sequence alignment of 3D-BLAST search results using the aminoglycoside 6′-NAT from *S.enterica* as the query. (**A**) The structural recurrences of five homologous proteins from the NAT superfamily are colored. Each structure is shown as a continuous color spectrum from red through orange, yellow, green and blue to violet. The PDB code and SCOP *sccs* are also indicated. (**B**) Sequence alignments of both structural alphabet and amino acid sequences of eight proteins. The location of secondary structures [i.e. strand (S) and helix (H)] and the motifs D, A and B are indicated. Motif A presents an invariant Arg-X-X-Gly-X-Gly/Ala segment (red columns) for four proteins (PDB code 1s3zA, 1ufhA, 1vhsA and 1bo4A) of the NAT family, and the secondary structures are obtained from DSSP based on the query protein. The structural alphabet sequences are more conserved than the amino acid sequences. The *E*-value between the query and homologous proteins from the 3D-BLAST results is given for each sequence.

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 5B). (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.

TIM barrel proteins

Thiamine phosphate pyrophosphorylase (PDB code 1 × i3), an α/β protein with a TIM barrel fold (27), catalyzes the formation of thiamine phosphate—an essential nutrient for humans (29). This protein is a structural genomics target for Southeast Collaboratory for Structural Genomics, which is a part of the Protein Structure Initiative (30). 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 α/β barrel and is by far the most common tertiary fold observed in protein crystal structures.

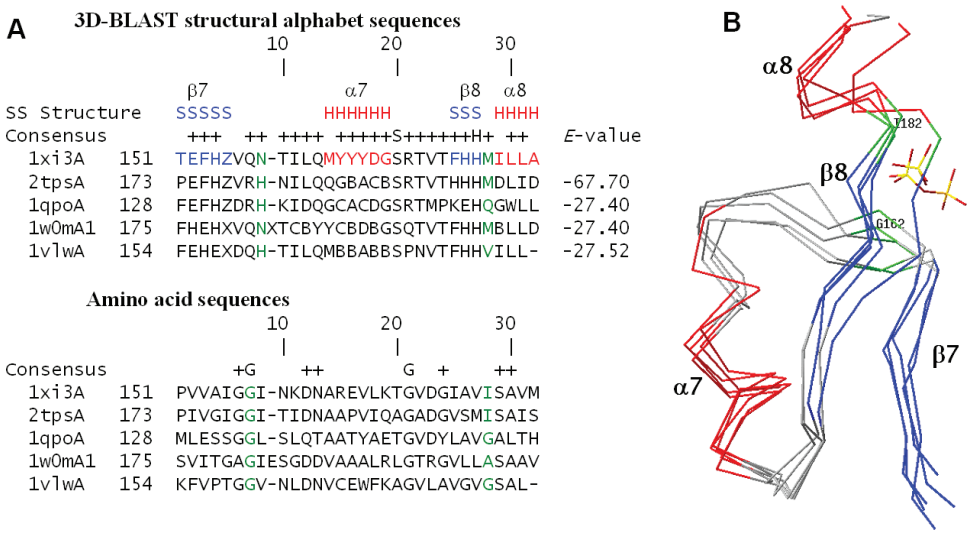


Figure 6. (**A**) Multiple sequence alignments and (**B**) multiple structure alignments resulting from a 3D-BLAST search using thiamine phosphate pyrophosphorylase from *P.furiosus* as the query. The aligned proteins are 1 × i3A and 2tpsA (thiamine phosphate synthase superfamily), 1qpoA (quinolinic acid phosphoribosyltransferase superfamily), 1w0mA (triosephosphate isomerase superfamily) and 1vlwA (aldolase superfamily). The alignment and superimposition region, a common phosphate-binding site of these five proteins, is from β7 to α8. The secondary structures of these proteins are indicated in red (helices) and blue (strands), whereas the remaining structures are shown in gray. The residues involved in phosphate-binding are indicated in green.

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.

When the *E*-value was restricted to 10^{-15}, 3D-BLAST identified 74 members from 16 SCOP superfamilies containing a TIM barrel fold (Table 4). Figure 6A and B show multiple sequence alignments and structure alignments, respectively, of five homologous proteins derived from the 3D-BLAST pairing alignments. These proteins, thiamine phosphate synthase (PDB code 1 × i3A 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 (31) 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.

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 6). 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*. (27)

3D-BLAST identified five homologous superfamilies, including quinolinic acid phosphoribosyltransferase, phosphoenolpyruvate and dihydropteroate synthetase-like. These distant relationships were also reported by Nagano *et al*. (27) using PSI-BLAST or IMPALA (32) with different iteration numbers. In addition, 3D-BLAST and sequential structure alignment program (SSAP) (33) 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.

DISCUSSION

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

Table 5. Average search time and performance of each program on 50 proteins selected from SCOP95-1.69

Program	Average time of a query (s)	Average time of a pair alignment (s)	Relative to 3D-BLAST	Correct assignment %	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 1.

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 1E. Therefore, 3D-BLAST is more sensitive when the query proteins are members of the 'all alpha' [e.g. PDB code 1v2z (34) and 1owa (35)]or 'all beta' [e.g. PDB code 1sq9 (36) and 1ri9 (37)] 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 (38) and 1q5f (39)]. In the third category, the query proteins had multiple domains [e.g. PDB code 1s35 (40) 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 (41), 1sq9 (36), 1ovy (42) and 1t3k(43)] 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.

Table 5 shows the average search time and average precision of 3D-BLAST, PSI-BLAST, YAKUSA (15), MAMMOTH (9) and CE (8) 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 (see Supplementary Table S1) from the set SCOP95-1.69 based on the lengths of these 516 query proteins. On average, 3D-BLAST required ~1.298 s 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 (6).

Among these five methods, MAMMOTH is the best and PSI-BLAST is the worst for these 50 queries (Table 5). 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 (44) (PDB code 1q9jA), Hypothetical protein Alr5027 (structural genomics target and PDB code 1vl7A) and avrpphf orf1 (45) (PDB code 1s28) (see Supplementary Table S1), 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 (46) (PDB code 1s6iA) and Putative mar1 (structural genomics target and PDB code 1 × 9gA), 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 (47) (PDB code 1ukvY) and 1s6iA (46).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

J.-M.Y. was supported by National Science Council and the University System at Taiwan-Veteran General Hospital Grant. The authors are grateful to both the hardware and software supports of the Structural Bioinformatics Core Facility at National Chiao Tung University. Funding to pay the Open Access publication charges for this article was provided by National Science Council.

Conflict of interest statement. None declared.

REFERENCES

1. Todd,A.E., Marsden,R.L., Thornton,J.M. and Orengo,C.A. (2005) Progress of structural genomics initiatives: an analysis of solved target structures. *J. Mol. Biol.*, **348**, 1235–1260.

2. Burley,S.K. and Bonanno,J.B. (2002) Structural genomics of proteins from conserved biochemical pathways and processes. *Curr. Opin. Struct. Biol.*, **12**, 383–391.
3. Burley,S.K., Almo,S.C., Bonanno,J.B., Capel,M., Chance,M.R., Gaasterland,T., Lin,D., Sali,A., Studier,F.W. and Swaminathan,S. (1999) Structural genomics: beyond the human genome project. *Nature Genet.*, **23**, 151–157.
4. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
5. Watson,J.D., Laskowski,R.A. and Thornton,J.M. (2005) Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.*, **15**, 275–284.
6. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
7. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
8. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
9. Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (Matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
10. Madej,T., Gibrat,J.F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins: Structure, Function, and Bioinformatics*, **23**, 356–369.
11. Aung,Z. and Tan,K.L. (2004) Rapid 3D protein structure database searching using information retrieval techniques. *Bioinformatics*, **20**, 1045–1052.
12. Shyu,C.R., Chi,P.H., Scott,G. and Xu,D. (2004) ProteinDBS: a real-time retrieval system for protein structure comparison. *Nucleic Acids Res.*, **32**, W572–W575.
13. Martin,A.C. (2000) The ups and downs of protein topology: rapid comparison of protein structure. *Protein Eng.*, **13**, 829–837.
14. Guyon,F., Camproux,A.C., Hochez,J. and Tuffery,P. (2004) SA-Search: a web tool for protein structure mining based on a structural alphabet. *Nucleic Acids Res.*, **32**, W545–W548.
15. Carpentier,M., Brouillet,S. and Pothier,J. (2005) YAKUSA: a fast structural database scanning method. *Proteins: Structure, Function and Genetics*, **61**, 137–151.
16. Levitt,M. (1992) Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.*, **226**, 507–533.
17. Bystroff,C. and Baker,D. (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.*, **281**, 565–577.
18. Kolodny,R., Koehl,P., Guibas,L. and Levitt,M. (2002) Small libraries of protein fragments model native protein structures accurately. *J. Mol. Biol.*, **323**, 297–307.
19. Takano,K., Yamagata,Y. and Yutani,K. (2000) Role of amino acid residues at turns in the conformational stability and folding of human lysozyme. *Biochemistry*, **39**, 8655–8665.
20. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
21. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
22. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
23. Pearl,F., Todd,A., Sillitoe,I., Dibley,M., Redfern,O., Lewis,T., Bennett,C., Marsden,R., Grant,A., Lee,D. *et al.* (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.*, **33**, D247–D251.
24. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
25. Hutchinson,E.G. and Thornton,J.M. (1996) PROMOTIF-a program to identify and analyze structural motifs in proteins. *Protein Sci.*, **5**, 212–220.
26. Vetting,M.W., Magnet,S., Nieves,E., Roderick,S.L. and Blanchard,J.S. (2004) A bacterial acetyltransferase capable of regioselective N-acetylation of antibiotics and histones. *Chem. Biol.*, **11**, 565–573.
27. Nagano,N., Orengo,C.A. and Thornton,J.M. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.*, **321**, 741–765.
28. Wolf,E., Vassilev,A., Makino,Y., Sali,A., Nakatani,Y. and Burley,S.K. (1998) Crystal structure of a GCN5-related N-acetyltransferase: *Serratia marcescens* aminoglycoside 3-N-acetyltransferase. *Cell*, **94**, 439–449.
29. Peapus,D.H., Chiu,H.J., Campobasso,N., Reddick,J.J., Begley,T.P. and Ealick,S.E. (2001) Structural characterization of the enzyme-substrate, enzyme-intermediate, and enzyme-product complexes of thiamin phosphate synthase. *Biochemistry*, **40**, 10103–10114.
30. Terwilliger,T.C. (2000) Structural genomics in North America. *Nature Struct. Biol.*, **7**, 935–939.
31. Wilmanns,M., Hyde,C.C., Davies,D.R., Kirschner,K. and Jansonius,J.N. (1991) Structural conservation in parallel b/a barrel enzymes that catalyze three sequential reactions in the pathway of tryptophan biosynthesis. *Biochemistry*, **30**, 9161–9169.
32. Schaffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
33. Orengo,C.A. and Taylor,W.R. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Meth. Enzymol.*, **266**, 617–635.
34. Uzumaki,T., Fujita,M., Nakatsu,T., Hayashi,F., Shibata,H., Itoh,N., Kato,H. and Ishiura,M. (2004) Crystal structure of the C-terminal clock-oscillator domain of the cyanobacterial KaiA protein. *Nature Struct. Mol. Biol.*, **11**, 623–631.
35. Park,S., Caffrey,M.S., Johnson,M.E. and Fung,L.W. (2003) Solution structural studies on human erythrocyte alpha-spectrin tetramerization site. *J. Biol. Chem.*, **278**, 21837–21844.
36. Madrona,A.Y. and Wilson,D.K. (2004) The structure of Ski8p, a protein regulating mRNA degradation: implications for WD protein structure. *Protein Sci.*, **13**, 1557–1565.
37. Heuer,K., Kofler,M., Langdon,G., Thiemke,K. and Freund,C. (2004) Structure of a helically extended SH3 domain of the T cell adapter protein ADAP. *Structure*, **12**, 603–610.
38. Mowat,C.G., Rothery,E., Miles,C.S., McIver,L., Doherty,M.K., Drewette,K., Taylor,P., Walkinshaw,M.D., Chapman,S.K. and Reid,G.A. (2004) Octaheme tetrathionate reductase is a respiratory enzyme with novel heme ligation. *Nature Struct. Mol. Biol.*, **11**, 1023–1024.
39. Xu,X.F., Tan,Y.W., Lam,L., Hackett,J., Zhang,M. and Mok,Y.K. (2004) NMR structure of a type IVb pilin from *Salmonella typhi* and its assembly into pilusstructural analysis of mutational effects. *J. Biol. Chem.*, **279**, 31599–31605.
40. Kusunoki,H., MacDonald,R.I. and Mondragon,A. (2004) Structural insights into the stability and flexibility of unusual erythroid spectrin repeats. *Structure*, **12**, 645–656.
41. Rubin,S.M., Pelton,J.G., Yokota,H., Kim,R. and Wemmer,D.E. (2003) Solution structure of a putative ribosome binding protein from *Mycoplasma pneumoniae* and comparison to a distant homolog. *J. Struct. Funct. Genomics*, **4**, 235–343.
42. Turner,C.F. and Moore,P.B. (2004) Purified NS2B/NS3 serine protease of dengue virus type 2 exhibits cofactor NS2B dependence for cleavage of substrates with dibasic amino acids *in vitro*. *J. Mol. Biol.*, **335**, 679–684.
43. Landrieu,I., da Costa,M., De Veylder,L., Dewitte,F., Vandepoele,K., Hassan,S., Wieruszeski,J.M., Corellou,F., Faure,J.D., Van Montagu,M. *et al.* (2004) A small CDC25 dual-specificity tyrosine-phosphatase isoform in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **101**, 13380–13385.
44. Buglino,J., Onwueme,K.C., Ferreras,J.A., Quadri,L.E. and Lima,C.D. (2004) Crystal structure of papA5, a phthiocerol dimycocerosyl

transferase from *Mycobacterium tuberculosis*. *J. Biol. Chem.*, **279**, 30634–30642.
45. Singer,A.U., Desveaux,D., Betts,L., Chang,J.H., Nimchuk,Z., Grant,S.R., Dangl,J.L. and Sondek,J. (2004) Crystal structures of the type iii effector protein avrpphf and its chaperone reveal residues required for plant pathogenesis. *Structure*, **12**, 1669–1681.
46. Weljie,A.M. and Vogel,H.J. (2004) Unexpected structure of the Ca2+-regulatory region from soybean calcium-dependent protein kinase-alpha. *J. Biol. Chem.*, **279**, 35494–35502.
47. Rak,A., Pylypenko,O., Durek,T., Watzke,A., Kushnir,S., Brunsveld,L., Waldmann,H., Goody,R.S. and Alexandrov,K. (2003) Structure of Rab GDP-dissociation inhibitor in complex with prenylated YPT1 GTPase. *Science*, **302**, 646–650.