國立交通大學

生物資訊所

碩士論文

PiSA-BLAST:快速蛋白質結構比對與資料庫搜尋工具

PiSA-BLAST: A New Tool for Protein Structure Alignment and Database Search

研 究 生：董其樺

指導教授：楊進木　教授

中華民國九十四年七月

PiSA-BLAST:快速蛋白質結構比對與資料庫搜尋工具

PiSA-BLAST: A New Tool for Protein Structure Alignment and Database Search

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

碩士論文

A Thesis Submitted to Institute of Bioinformatics

National Chiao Tung University in partial Fulfillment of the Requirements

for the Degree of Master in

Bioinformatics

July 2005

Hsinchu, Taiwan, Republic of China

中華民國九十四年七月

PiSA-BLAST:快速蛋白質結構比對與資料庫搜尋工具

學生：董其樺　　　　　　　　　　　　　　　　指導教授：楊進木

國立交通大學生物資訊所碩士班

摘　　要

近年來隨著蛋白質結構數量快速成長，有效搜尋結構資料庫的方法愈形重要。當一個新的蛋白質結晶產生後，研究者會希望得知該蛋白質是否跟其他已知結構的蛋白質相似，以及其相似程度。由於蛋白質結晶結構的數量龐大，研究者便十分需要一個準確而有效率的搜尋相似結構之工具。在本研究論文中，我們發展一套新的工具「PiSA-BLAST」，除了提出準確的比對結果外，也能大幅提昇結構搜尋的執行速度。

這套工具依據 DSSP 程式所定義的蛋白質特殊資訊：kappa 角與 alpha 角，利用分群演算法加以分析後得一轉換規則表。依據此規則表，將蛋白質結構資料庫裡所有已知結構的蛋白質轉換成一級序列，並建成序列資料庫。根據此序列資料庫，我們同時也發展一套新的計分陣列，將之用來計算序列比對時的比對分數。接著，我們結合知名的序列比對工具「BLAST」，在輸入一欲查詢、比對的蛋白質結構後，不需真正疊合兩個三級結構，即能快速地從含有大量序列的結構資料庫搜尋、比對，最後能獲得相似蛋白質的清單。

我們從 SCOP 及 PDB 資料庫中挑選出五套測試資料，以驗證 PiSA-BLAST 之效能。我們以 108 個查詢結構(query structures)在 SCOP 95 的搜尋結果為例，此資料庫包含 9,354 個蛋白質結構， PiSA-BLAST 及 CE 在 108 個查詢的平均準確度分別為 78.2%與 82.1%，PiSA-BLAST 總搜尋時間只需 34 秒，遠快於 CE 搜尋所需的 1,169,832 秒。另外，PSI-BLAST 的平均準確度則為 69.8%，並共花費 18.3 秒。根據本篇論文的研究結果，顯示下列結論：一、PiSA-BLAST 能以接近 BLAST 的速度搜尋結構資料庫，並較 CE 快上 34,000 倍左右。二、PiSA-BLAST 能獲得接近 CE 的準確度，同時較以胺基酸為基礎的序列比對工具，如 BLAST、PSI-BLAST 等，提供更精確的搜尋結果。這些結果顯示，在結構比對時，我們所發展的結構編碼以及計分陣列確實正確、可用。三、如同 BLAST 在執行序列比對時能輸出一 e-value，PiSA-BLAST 亦可在搜尋結構時提供此輸出值。經測試，當 e-value 小於閾值 e^{-15} 時，PiSA-BLAST 可達到 90%的準確度。四、PiSA-BLAST 可成為一個結構比對的快速篩選工具，先執行一次快速比對，輸出多個結果後再利用其他速度較慢，但比對方法詳盡、可信的工具如 CE、DALI，作第二次的分析。五、PiSA-BLAST 已建立網頁服務，使用者能在線上即時搜尋結構資料庫。綜合以上所述，本研究□　愀Ĥ ▯ 因體學與蛋白質體學應有相當的貢獻。

PiSA-BLAST: A New Tool for **P**rote**i**n **S**tructure **A**lignment and Database Search

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

Institute of Bioinformatics
National Chiao Tung University

ABSTRACT

The structural database searching has become increasingly important with growing numbers of known protein structures. This increase was near exponential in the early 1990s and has become linear over the past several years. As more and more the availability of the growing number of protein crystal structures, the demand for a very fast and accurate method to searching for structures similar to a query structure is high. In this thesis, we have developed a new tool, termed PiSA-BLAST for protein structure database search that does not require the alignment of two 3D structures.

Here we have developed a new method for the protein structure alignment by transforming 3D structures into 1D sequences. This method use the information of kappa and alpha angles, derived from DSSP program, to represent the protein 3D structure. Based on the segment information and clustering method, we transform the structural information with kappa and alpha angles into coded regions. After that, each protein with 3D structure is able to transfer into 1D sequence and we could develop a new substitution matrix that can be used as the scoring matrix of sequence alignment for 23 new codes. These encoded sequences are collected as a structure database. Launching BLAST, a well-known sequence alignment tool, to search structure database in a short time and we will get a list of proteins that are similar in structure.

We evaluated PiSA-BLAST on five diverse data sets from SCOP and protein data bank. For the dataset SCOP 95 with 108 queries on 9,354 protein domains, the average precisions of PiSA-BLAST and CE are 78.2% and 82.1%, respectively, and the total executing times are 34 seconds for PiSA-BLAST and about 1,169,832 seconds for CE. The average precision is 69.8% and time is 18.3 seconds for PSI-BLAST. Based on these experiments, we summarized several observations: (1) PiSA-BLAST is as fast as BLAST for protein structure database search and is 34,000 times faster than CE on the database SCOP 95. (2) The accuracy of PiSA-BLAST closes the accuracy of CE and much better than BLAST and PSI-BLAST which are based on amino-acid sequences. These results imply that our structural new codes and substitute matrix are useful for protein structure alignment. (3) PiSA-BLAST is able to provide a significant e-value with e^{-15} for structure database search as the e-value with e^{-3} in BLAST for sequence database search. PiSA-BLAST achieved about 90% accuracy for a query when e-value is less than e^{-15}. (4) PiSA-BLAST is a useful filtering tool before performing a detailed database search, such as CE and DALI. (5) PiSA-BLAST is able to provide real-time web services for protein structure database search as BLAST in protein sequence search. We believe that this issue is important for structural genomics and proteomics.

誌　　　謝

在這三年來的研究所生涯中，首先衷心感謝指導教授—楊進木老師。在整個研究過程中，老師花費了很多心力及時間，悉心指導我的論文與研究。老師對於學術研究的細心、嚴謹、堅持與執著，將是我未來持續學習的目標。

這本論文能夠順利完成，要感謝實驗室所有同學的幫忙。特別感謝章維，他早期的初步研究以及撰寫程式的功力，都為我奠定了穩固的基礎，讓我的研究得以順利進行。

另外要感謝我的兩位好友—開文及靜婷。在我心情最低落的時候，給予我無盡的鼓勵、安慰與溫暖。在我情緒最愉快的時刻，陪我一起出遊、聊天，抒解壓力。沒有他們的陪伴，我很難走到現在。

最後謝謝我的家人，默默地在背後給我支持、鼓勵與動力，讓我可在求學過程中無所顧忌，全力衝刺。

要感謝的人太多太多了，一切就感謝上天吧。

其樺

夏'05

CONTENTS

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivations and Purposes

Protein structures are being determined at a very rapid rate; as of 07-Jun-2005, there were more than 31000 proteins in the Protein Data Bank (PDB) and the number is increasing daily and rapidly. As a result, faster tools for structural comparison and database searching become essential. Protein structure comparisons have been made since the very early days of protein crystallography. These pioneering early works have been reviewed [1]. However, these early methods are too slow to handle the volume of data that is now available.

In general, we cannot detect the similarity of two remotely homologous proteins by sequence comparison alone because comparing the amino acid sequences of the proteins cannot provide sufficient information required by the biologist. Therefore, we need to compare their 3D structures in order to determine their similarity as the 3D structures are better preserved than the sequences throughout the evolution. We usually compare a protein structure against a database of other protein structures to find the structures that are similar to it.

Here we develop a novel structure alignment tool, termed PiSA-BLAST, for protein structure comparison and fast database searching. PiSA-BLAST cannot only scan whole protein database as fast as sequence alignment but also obtain acceptable accuracy. Our method use segment information such as kappa and alpha angles, derived form DSSP program, to represent the local 3D structures of proteins. With nearest neighbor clustering algorithm [2], we transform the 2D information of kappa and alpha angles into 23 new coded residues. By this way, each protein 3D structure in PDB could be described as a 1D sequence. After transforming, we develop a new substitution matrix for 23 codes and replace default matrix of sequence alignment with the new one. The structure comparison is established by a well-known sequence alignment program such as BLAST [3, 4] to search for similar coded sequences that are converted from other protein. Our results show that PiSA-BLAST is 5000 times faster than the popular CE method for structural database searching, while its overall accuracy is only slightly inferior to that of CE. Although our new methods could not provide the same accuracy as the results of CE, it can be used as a pre-filtering tool before performing a detailed database search by other more delicate but slower structure alignment tools.

1.2 Related Works

As in past research, the different amino acid sequences may determine similar protein structures [5, 6]. If there is 30% or above sequence identity between two proteins, these two proteins may have quite similar 3D-structure [7]. However, sequence comparison alone cannot provide required information in the twilight zone of protein sequence alignments [8]. If only using sequence alignment to detect protein structure similarity, it will lose some proteins which are with low sequence identity and high structure similarity. Structural comparison must be performed in this case.

Many methods have been proposed and implemented for structural comparison. The classical pairwise comparison methods include DALI [9], VAST [10, 11] and CE [12]. These are the two-level methods, which start with finding the matching pairs of secondary structure elements (SSEs) or $C\alpha$ backbone fragments, and then go into the detailed finding of the matching Cα atom pairs. The distance matrix alignment (DALI) algorithm is the core of FSSP [9]. This algorithm is based on building residue-to-residue distance matrices and using Monte Carlo to optimize distance matrix comparing. The vector alignment search tool (VAST) define protein secondary structure elements as vectors to compare 3D protein structures and determine the protein structure neighbors [10, 11]. In the method of combinatorial extension (CE), aligned fragment pairs are divided in a protein. After that, these pairs are joined into an optimal path for the full alignment [12]. These methods can provide us with the good quality answers. But when performing a database search, they all have to use exhaustive searching, which results in slow response times.

TopScan [13] are examples of pairwise comparison methods that take SSEs as basic elements to be compared. These methods are less accurate, but much faster than the two-level methods. However, when searching against a very large database, these methods still cannot provide the required quick response time. The design strategy of ProtDex2 [14] is to apply the IR approaches using SSEs as the basic elements in order to perform rapid database searching without having access to every 3D structure in the database. ProtDex2 first build an inverted-file index based on the feature vectors of the relationships among the SSEs from all the protein structures in the database.

Unlike 1-dimensional sequence comparison, structure alignment is much more complex and computationally expensive to compare two structures to determine their similarity. Although some of the related works are very efficient for pair-wise structure comparison, the main disadvantage of these methods is that they practice exhaustive searching to compare the query structure against all protein structures in the database when performing a structural database search. Exhaustive searching can give a satisfactory response time until today.

However, giving the rapid growth rates of the structural databases in the near future, such a structural database searching will be restrictedly expensive to be performed.

With a query protein structure, we search through the database and report the structures that are similar to the query structure. There may define a similarity threshold, and the structures whose scores are equal to or above the threshold are reported. Because the execution time of global searching through a structural database is very expensive, some fast but rough searching methods such as TopScan [13] and ProtDex2 [14] can be used as a pre-filter before performing the further database searching. In this way, the structures that are very improbable to be included in the report could be eliminated after a quick screening before going into the expensive comparison.

1.3 Thesis Overview

We develop a novel sequence-based structure alignment: PiSA-BLAST for fast database searching. In chapter 2, we have prepared training set from ASTRAL SCOP database 1.65 40% set. We divide domain proteins of training set into many segments that have various kappa and alpha angle. Then, we find representative segments of each kappa and alpha angle cell and use cluster algorithm to group these representative segments. After that, we assign a new code for each representative group. Next, we need to develop a substitution matrix for new codes and use it to replace default matrix for sequence alignment tool. Finally, we can run sequence alignment tool to do fast protein structure searching in database.

In chapter 3, we demonstrated the conformation of representative segments that are belonging to the same coding region and the new substitution matrix for representative segments. In addition, we evaluated the database searching time and screening performance of

PiSA-BLAST with several testing sets by precision, recall, false positive rate and ROC curve. Besides, we discussed the relationship between precision, sequence identity, structure similarity and theoretically expected number and given some examples to explain PiSA-BLAST how to works on practical applications and what weakness it has in this chapter.

Chapter 4 presented some conclusions and future perspectives. Our major contribution is to develop a novel fast structure alignment tool for protein database searching. The coded sequence has biological meanings. From 3D to 1D level, PiSA-BLAST can decrease execute time by translating 3D-structure to 1D-sequence and using sequence level to align structure. From 1D to 3D, PiSA-BLAST can enhance the accuracy of sequence alignment for structure searching by adding segment information into 1D-sequence. Because of fast structure database searching, we can apply PiSA-BLAST in biological issues like fold assignment and homology searching. Furthermore, PiSA-BLAST can be used on several practical applications, for example, multiple structure alignment, finding structure motifs, protein function prediction, and protein-protein interaction in the future.

Chapter 2

Materials and Methods

Step-by-step illustration of the PiSA-BLAST methodology is showed in Figure 1. Given one known 3D structure for query protein in a structure database. Every 3D structure in database can be divided into 5-mer structure segments by its kappa and alpha angle. After determining segments, we translate these segments into encoded sequence according kappa and alpha clustering map. The following step is to run structure alignment with encoded sequence using sequence alignment tool: BLAST. As the result, we can gain alignment score, structure similarity and even superposition sites of two aligned protein.

The flowchart of research step is shown in Figure 2. First, we prepare training set from ASTRAL SCOP database 1.65 40% set [15, 16]. Second, we divide domain proteins of training set into many segments that are have various kappa and alpha angle. Then, we find representative segments of each kappa and alpha angle and use cluster algorithm [2] to group these representative segments. After that, we assign a new code for each representative group. Next, we need to develop a substitution matrix for new codes and use it to replace default matrix for sequence alignment tool. We can use sequence alignment tool to do fast protein structure searching in database and evaluate the performance. Finally, we apply the PiSA-BLAST on practical application.

2.1 Preparing Training Set from Protein Structure Database

We prepare 792 pairs domain proteins in ASTRAL SCOP database 1.65 40% set [15, 16] for developing of 3D-1D coding and establishing new substitution matrix. The principle of training set collecting is as follows.

First, we select families with at least two domain proteins and totally choice 882 families. In these families, select one pair domain per ten domain proteins in random. Each pair domain belongs to the same family and sequence identity of each pair domain is less than 40%. Second, after structure alignment of CE, the RMSD in pair domain proteins is less than 5Å. Third, the residues in all selected domain proteins are exclude “X”.

We expect that our training set can reflect the real condition in composition of amino acids. Figure 3 shows that Comparison the amino acids compositions of our train set, including 1584 proteins for encoding the structured codes and the substitute matrix, with three well-known structure databases (DSSP database [17], SCOP 95 and SCOP 40 database [15]). The distributions of amino acids compositions of these four databases are similar. So, our training set can provide right and meaningful information.

2.2 Dividing Protein Structures into Segments by Kappa-Alpha Angle Map

The kappa angle is described as virtual bond angle (bend angle) defined by the three C-alpha atoms of residues I-2, I, I+2. The range of kappa angle is 0° to 180°. The alpha angle is described as virtual torsion angle (dihedral angle) defined by the four C-alpha atoms of residues I-1, I, I+1, I+2. The range of alpha angle is –180° to 180° (described at http://www.cmbi.kun.nl/gv/dssp/de-scrip.html#SECSTRUC). According to the definition of kappa and alpha angle, we define the local structure with 5 residues long as a segment.

792 domain protein pairs have been divided into total 263696 segments. These segments

are separated by various kappa and alpha angle. Figure 4 shows the distribution of 263696 segments in various kappa and alpha angle. The color bar on the right side shows the distribution scale. These segments are encoded into 23 codes based on the distributions of kappa and alpha angle. The helix-like segments (e.g., A, B, C and D) have more than 9000 segments whose alpha angle ranging from 40° to 60° and kappa angle ranging from 100° to 120°. The strand-like segments (e.g., E and F) have over 3000 segments with alpha angle ranging from -180° to -140° and kappa angle ranging from 0° to 20°.

Because of the large number of segments, we need to cluster these segments for representative segments deciding and meaningful codes assigning.

2.3 Finding Representative Segments and Using Nearest Neighbor Clustering Algorithm for New Codes Assigning

There are total 648 cells on kappa and alpha angle map *K*. Each cell includes many segments shown in Figure 4. We use the simple way as follows to decide one representative segment for each cell.

We building inter-segment distance matrix for one cell. Let d_{ij} be the structure distance (measured by superimpose program [18]) between segment *i* and segment *j*. The number of *i* and *j* is equal to the number of segments for this cell. Then, we summarize each column of the distance matrix and get the minimum of sum of column. Hence we select the representative segment for one cell depend on its lowest total structure distance among other segments.

After finding representative segment for every cell, we use nearest neighbor clustering algorithm [2] to group these representative segments with similar conformation. The algorithm is based on calculating a matrix, *D*, where *N* is the number of representative segments to be clustered. The matrix *D* is stored with the values of Rmsd for inter-representative segments. D_{ij} is a measure of structure similarity (computed by superimpose program [18]) between representative segments *i* and *j*. Clusters are formed recursively by adding other representative segments according to the nearest neighbor criterion. The method of nearest neighbor clustering is as follows:

Input:

(1) The matrix *D* is stored with the values of RMSD for all inter-representative segments. D_{ij} is a measure of structure similarity between representative segments *i* and *j* ($0 \leqq i, j \leqq 648$).

(2) The matrix *K* is collected with the numbers of segments with various kappa and alpha angle. K_{ab} is a number, which means how many segments in alpha angle a° and kappa angle b° ($0 \leqq a \leqq 36$, $0 \leqq b \leqq 18$).

Output:

The encoding rule map *E* point out that each cell with various alpha-kappa angle could be assign one letters of the alphabet. The size of encoding rule map is 36*18 according the range of kappa and alpha angle. The range of alpha angle is observed into 10° interval ranging from -180° to 180°. The range of kappa angle is observed into 10° interval ranging from 0° to 180°.

Step:

(1) Select one cell of *E* with particular kappa-alpha angle which the K_{ab} is the most and this cell E_{ab} did not assign any code yet to be the center of a cluster.

(2) Assume that the representative segment of this center is representative segment *i*.Sort the value from $D_{i,0}$ to $D_{i,648}$.

(3) According the result of sorting, from top to bottom, group every cell $E_{a'b'}$ repeatedly into the cluster with center E_{ab} if the $E_{a'b'}$ fit in with following conditions.

(3.1) Given a threshold, *t*, on the nearest neighbor distance. Assume that the representative segment of $E_{a'b'}$ is representative segment *j*. The D_{ij} is less than *t*.

(3.2) Given a threshold, *u*, for the maximum fragments number. If group into the cluster, the summation of the number of fragments in this cluster is still less than *u*.

(4) Check if this cell $E_{a'b'}$ has already grouped to other cluster.

(4.1) If not, group the cell $E_{a'b'}$ into the cluster with center E_{ab} and record the D_{ij} for the optimized clustering.

(4.2) Otherwise, compare the value of the previous and present record of D_{ij}.

(4.2.1) If the present record of D_{ij} is less than the previous one, $E_{a'b'}$ would be re-assign into the present cluster. However, the sum of the number of fragments in this cluster must be less than *u*.

(4.2.2) If the previous record of D_{ij} is less than the present one, do nothing and keep previous cluster.

(5) Repeat step 1 to 4 until every cells of *E* is clustered to 21 groups.

(6) First group has only one cell of *E*. This cell E_{ab} is assigned to code "A". The code "A" with alpha angle more than 46° and kappa angle less than 114° will be assigned to another code "Y".

(7) Every cells of E in second group are assigned to code "B", ones in third group are assigned to code "C", and etc. Ones in last group are assigned to code "X". There are exclude J, O and U in code assignment.

(8) If the K_{ab} is less than 40, this cell would be assigned to code "Z".

(9) Every E_{ab} is assigned to one code and output result of encode rule map E.

Here, the threshold, t and u, is given depending on how many groups we want. Here the threshold t is 0.72, u is 18450, and 21 groups is made. The threshold u is given by the 7% of the number of total segments.

Each group in various cells is assigned to a new code. There are 21 codes named as letter “A” to “X” (exclude “J”, “O” and “U”). If the number of segments in one cell is less than 50, this cell will be assigned to Code “Z”. In addition, when the structure is coding to sequence, the new code “A” with alpha angle more than 46° and kappa angle less than 114° will be assigned to another code ”Y”.

2.4 Generating a Substitution Matrix for 23 New Codes

The method of generating a substitution matrix refer to BLOSUM62 [19]. The elements of the substitution matrix are calculated as follows. For each residue position in the training set of pair database of aligned structural pairs, the statistics is counted at each aligned position. Each protein chain is considered to be a coded sequence aligned to a structure. The substitution score for coded sequence i and j with homologous structure is given by the information value [20].

Let the total number of amino acid i, j pairs $(1 \leqq j \leqq i \leqq 20)$ for each entry of the frequency table be fij. Then the observed probability of occurrence for each i, j pair is

$$q_{ij} = f_{ij} / \sum_{i=1}^{20} \sum_{j=1}^{i} f_{ij} \qquad (1)$$

Next we estimate the expected probability of occurrence for each *i, j* pair. It is assumed that the observed pair frequencies are those of the population. In general, the probability of occurrence of the ith amino acid in an *i, j* pair is

$$p_i = q_{ii} + \sum_{j \neq i} q_{ij} / 2 \qquad (2)$$

The expected probability of occurrence eij for each *i, j* pair is then *pipj* for *i* = *j* and *pipj* + *pjpi* = 2 *pipj* for *i* ≠*j*.

$$e_{ij} = \begin{cases} p_i p_j & if \quad i = j \\ 2 p_i p_j & if \quad i \neq j \end{cases} \qquad (3)$$

Then, the substitution matrix scores are then defined as

$$s_{ij} = \lambda \log_2 \frac{q_{ij}}{e_{ij}} \qquad (4)$$

where λ is an arbitrary positive rational number. Here, λ is given 1.89 for the best performance and efficiency.

The following describes the overall procedure for generating the λ value and optimized gap penalties. In the first step, we tested the λ value observed into 0.5 interval ranging from 1.0 to 10.0. The result revealed that the λ value between 1.5 and 2.5 is better. The second step is verifying the detail λ value observed into 0.1 interval ranging from 1.5 to 2.5. Furthermore, we test the six sets of open and extend gap penalty and λ value to find out the optimized parameter for the performance of PiSA-BLAST. As the Figure 9 showing, the best combination of parameters is 8 for open gap penalty, 2 for extend gap penalty, and from 1.8 to 1.9 for the λ value. Finally, we experimented the best λ value from 1.82 to 1.93 according to the observation of results in second step. Figure 10 demonstrates that we acquire the best performance of database searching when λ value is 1.89.

2.5 Structure Searching by Sequence Alignment tool: BLAST and PSI-BLAST

We download standalone BLAST 2.2.10 [3, 4] from:

ftp://ftp.ncbi.nlm.nih.gov/blast/executables/snapshot/2004-12-05/

The default matrix: "BLOSUM 62" is replaced by the substitution matrix for 23 new codes. Use program: "formatdb" to create our own database made of 3D-1D coded and FASTA formatted protein sequences for BLAST searching. We execute BLAST by program "Blastall". Blastall may be used to perform all five flavors of blast comparison. A typical use of blastall would be to perform a "blastp" search (protein vs. protein) of a query file called INPUT would be:

```
blastall -p blastp –d DATABASE –i INPUT–o OUTPUT –M BLOSUM62 -G 8 -E 2 -F F
```

The output is placed into the result file OUTPUT and the search is performed against the 'DATABASE' database. Other blastall options showed above are "-M BLOSUM62" which is default scoring matrix, "-G 8 –E 2" which means that open gap penalty is 8 and extend one is 2, and "-F F" that is to tell blastall do not filter query sequence.

Furthermore, we also combine Position-Specific Iterated BLAST, or PSI-BLAST, with our method for detail database searching [3]. The PSI-BLAST program can do an iterative search in which sequences found in one round of searching are used to build a score model for the next round of searching. When the PSI-BLAST is producing, the position-specific matrix for round i+1 is built from a constrained multiple alignment among the query and the sequences found with sufficiently low e-value in round i.

There is another command to perform PSI-BLAST.

blastpgp -d DATABASE -i INPUT -o OUTPUT -F F -G 8 -E 2 -j 3 -t F -h 1e-15

Program "blastpgp" takes a protein query and perform PSI-BLAST search to create a position specific matrix using a protein database. Some of arguments used in PSI-BLAST are the same as BLAST. There are different options between BLAST and PSI-BLAST, such as "-j 3" which is the maximum number of rounds, "-t F" which means that program do not use composition based statistics, and "-h 1e-15" that is the e-value threshold for including sequences in the score matrix model. The e-value threshold is 0.001 in default. However, in order to obtain correct result and best performance, we change the value from 0.001 to 1e-15 for PiSA-BLAST.

The top part of the output of PSI-BLAST for each round distinguishes the sequences into: sequences found previously and used in the score model, and sequences not used in the score model. The output currently includes lots of diagnostics requested by users at NCBI. To skip quickly from the output of one round to the next, search for the string "producing", which is part of the header for each round and likely does not appear elsewhere in the output. PSI-BLAST "converges" and stops if all sequences found at round i+1 below the e-value threshold were already in the model at the beginning of the round.

2.6 Evaluating the Performance

We compare both the results from BLAST, PSI-BLAST, CE and PiSA-BLAST against SCOP classifications [15, 16] which is regarded as the golden standard by the biologists. SCOP classification hierarchy is made of 4 levels: class, fold, superfamily and family among which family is the most detailed classification. In our test, if a protein in the result set belong to the same SCOP family as the query protein, it is counted as a true hit.

We used 3 testing sets for evaluating the performance. First two experiments are refer to Aung *et al.* [14]. One involved a small database and a limited number of queries, and the other involved a large database and a greater number of queries. Third experiment involved the same number of queries as second testing set and SCOP 1.65 95% database.

In first experiment set, there are 10 proteins from Globins family (a.1.1.2 in SCOP) and 10 proteins from Serine/Threonin kinases family (d.144.1.1 in SCOP) from the representative ASTRAL dataset with less than 40% sequence homology. These 20 proteins are designated as the query proteins. Table 1 shows that the small test set selected from previous work [14]. There are 200 members in the database and 20 queries in two SCOP families are listed. In addition, other 180 proteins were selected from four major classes (All-α, All-β, α/β and $\alpha+\beta$) of the same representative dataset. These 180 proteins were combine with the above-mentioned 20 query proteins to form the small target database of 200 proteins.

In second and third testing sets, we conduct another experiment using a large database containing 33311 proteins which is refer to Aung *et al.* [14] and containing 9354 domain proteins in SCOP 1.65 95% database. From them, Zeyar select 108 query proteins which belongs to 108 medium-size families (with $\geq$40 and $\leq$180 members) from four major classes, and which have less than 40% sequence homology to each other. The lists of 108 query proteins are given respectively in Table 2.

Four common metrics were used to evaluate the quality of database searching, including precision, recall, false positive rate and 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 $(T_h\text{-}A_h)/(T\text{-}A)$, respectively. Here, A_h is the number of true hits in the hit list, T_h is the total number of domain proteins in the hit list, A is total number of true hits in the databases, and T is 33311 or 9354, the total number of domain proteins in these two large databases. The ROC curve plots the sensitivity against the "1-specificity". The sensitivity is equal to recall, and the "1-specidicity" is equal to false positive rate.

True hit, also called a relevant retrieval, is defined as an event of retrieving a protein from the database that belongs to the same "family" as the query. BLAST, PSI-BLAST and PiSA-BLAST can retrieve subject proteins by e-value and alignment score. However, CE did not provide sorted retrieval list when we use CE to perform one-against-all searching. For this reason, we need to sort the searching results by ourselves in order to obtain the retrieval list. In Figure 12, Recall-precision curves of CE using z-score and rmsd to order searching results on 108 queries searching the SCOP 95 database. The results of CE searching which are sorted by z-score are much more accurate than by rmsd. Hence, the results of CE searching are sorted by its Z-score.

We test the database searching time for CE, BLAST, PSI-BLAST and PiSA-BLAST on the same machine (LINUX platform with Pentium IV processors 2.8GHz and 2GB memory). We use the default parameters for CE, BLAST, PSI-BLAST and 7 target databases: small database including 200 proteins, large database including 33311 proteins, PDB, nr-PDB, SCOP 1.65 database, SCOP 1.65 95% and SCOP 1.65 40% as the database searching for both methods.

2.7 Practical Applications

The advancements in the protein crystallography to determine the structures of the protein molecules, the sizes of the structure databases such as PDB are growing at a very fast rate. It is possible that many new protein structures have been crystallized but their function and fold is still unknown.

Because of fast structure database searching, we can apply PiSA-BLAST in biological issues like fold assignment and homology searching. Here, we use PiSA-BLAST on biological application: fold assugnment and function predition.

We took 108 proteins that is the same as above testing set as the query to perform PiSA-BLAST database searching. These proteins is well-known function and have been assigned to particular fold family at SCOP and CATH database.The search was performed against the PDB database which is published at 19 April in this year. Then, we observed the top rank 100 proteins at the output of PiSA-BLAST.

We assume that there are several proteins which is unknown fold or function in top rank 100. If these new proteins are certainly similar to the query protein according its high statistical significance, we could predict these their function and fold family confidently.

Chapter 3

Results and Discussions

3.1 Representative Segments of 23 New Codes

The representative segments and 23 new codes defined by nearest neighbor clustering method are meaningful. Figure 4 also shows the result of the new code with 23 letters of alphabet mapping into kappa and alpha angle with the distribution of segments. Figure 5 shows the accumulated distributions of 20 kinds of amino acids and 23 structured codes in training set. The accumulated distribution of 23 codes is similar to the distribution of 20 amino acids. The most number in 20 amino acids is amino acid, leucine (L), and the ratio is 9.26%. The most quantity in 23 new codes for PiSA-BLAST is H and the ratio is 6.99%.

Figure 6 indicates the conformations of the representative segments of 23 new codes. The representative segments at code A, Y, B, C and D are called helix segment and segments at code G, I, L are called helix-like segment according to its conformation and distribution of DSSP secondary structure. The representative segments at code E, F and H are called strand segment and segments at code K, N are called strand-like segment. Representative segments at other codes are classified into loop-like segment and display different conformations between helix-like and strand-like segments. Figure 7 is another evidence to display the conformations of representative segment in each cell of four main groups. As the conformations show, it is clear to see that the structure of segments is very similar in same secondary structure defined region.

Figure 8 demonstrates that the distribution relationship between 23 new codes (in PiSA-BLAST) and 8 secondary structure codes (in DSSP [17]). It is clear to illustrate that the distribution of helix, helix-like, strand and strand-like segments defined by PiSA-BLAST are high related to secondary structures in DSSP and explain why the conformation of representative segments is similar in same coding. As shown in Figure 8(A), helix and helix-like segments: "AYBCDGIL" have large number in helix codes: "HGI" which is defined by DSSP. In the Figure 8(B), we also see strand and strand-like codes: "EFHKN" defined by PiSA-BLAST have quite a few of distribution in DSSP strand code: E and B.

According conformations in Figure 7 and the distribution of secondary structure in Figure 8, we can prove 23 codes in the encoding rule map are meaningful.

3.2 The Substitution Matrix for 23 New Codes

The substitution matrix of 23 new codes is given in Figure 11. The matrix offers insights about substitution preferences of 23 new codes between homologous structures. All identical new codes having the same secondary structure have positive substitution scores. The scores on the diagonal cells are much higher than the scores on the non-diagonal cells. Red dot-square part (A, Y, C, B, and D) is the scores of aligning helix codes to helix codes and blue dot-square part (H, E, and F) is the scores of aligning strand codes to strand codes. The scores of aligning helix codes to strand codes are the smallest.

In Figure 11, red dot-square is shown as the substitution scores of helix and helix-like codes. The mean of scores between helix and helix-like codes is greater than zero. Blue dot-square is shown as the substitution scores of strand and strand-like codes. The average of these scores in the blue square is greater than zero, too. In the yellow region, it is display positive score on the substitution matrix. In addition, orange region shows that there are negative substitution scores in the matrix between helix and strand codes, which are dissimilar secondary structures. Further more, light yellow region shows clearly that there are smaller substitution scores than ones in yellow region between helix and helix-like codes or strand and strand-like codes.

The above relationships are well known, showing that the substitution matrix embodies conventional knowledge about structure information in proteins.

3.3 Evaluating Statistical Significance

PiSA-BLAST is more accurate than BLAST and other tools for structure database searching. As shown in Table 3, we compare PiSA-BLAST with well-known tools for small database searching. In the Table 3, row i represents the ranking under the various methods to retrieve i relevant answers. For example, row 6 says that when 6 answers are required, the top 6 ranked answers from DALI, CE, ProtDex2 and PiSA-BLAST are the 6 relevant answers from the same family as the query; while BLAST ranks the 6 relevant answers among the top 18 retrievals.

We can see that PiSA-BLAST appears the good performance as good as CE and DALI in small database searching. In order to obtain all the relevant answers, PiSA-BLAST retrieves same number of proteins as the detailed comparison methods of DALI and CE. BLAST and PSI-BLAST using amino acid sequence to search homologous proteins have to retrieve more proteins than DALI, CE and PiSA-BLAST using structural information to search database.

The accuracy comparison is shown in Figures 13 and 14. The results are shown as recall-precision curves. Again, a relevant retrieval is defined as an event of retrieving a protein from the database that belongs to the same 'family' as the query. In Figure 13, the recall-precision curves of five alignment tools for 108 queries on the large database of 33311 proteins indicated in Table 2 is given. It shows clearly that PiSA-BLAST is the best and TopScan is the worst among these five approaches. BLAST and PSI-BLAST using sequence information only cannot provide right relevant retrieval, even PSI-BLAST search repeatedly. The results of ProtDex2 and TopScan, two fast structure alignment tools, are summarized from [14]. ProtDex2 [14] and TopScan [13] can search database quickly on sequence level but lost quite a few structural information.

In Figure 14, we compare the performance of PiSA-BLAST with CE, PiSA-PSI-BLAST, BLAST and PSI-BLAST methods on SCOP 1.65 95% database. Recall-precision curves in Figure 14 show obviously that CE supplies the more accurate than other methods. The accuracy of PiSA-BLAST closes the results of CE and PiSA-BLAST is about 34000 times fast than CE. Besides, PiSA-PSI-BLAST surprisingly only slightly improves PiSA-BLAST. In contrast, the performance of PSI-BLAST is much better than BLAST. At 10% recall, the precision of BLAST and PSI-BLAST is the same high as PiSA-BLAST. At 20% recall, PiSA-BLAST and PiSA-PSI-BLAST can supply the same accuracy as CE. However, when the recall is 20% and above, the precision of BLAST and PSI-BLAST decrease quickly.

The results of ROC curve for 108 queries on large databases searching are shown in Figures 15 and 16. PiSA-BLAST and PiSA-PSI-BLAST can appear the performance close to CE and are more accurate than sequence alignment tools, BLAST and PSIBLAST. Table 7 shows that the average precision of BLAST, PSI-BLAST, PiSA-BLAST, CE and PiSA-PSI-BLAST in SCOP95% database searching with each query protein.

We discuss the result of CE and PiSA-PSI-BLAST as following description. The overall accuracy of CE is better than other methods. However, the results of homology searching of CE may show weakness and even worse than PiSA-BLAST in some queries. As shown in Table 7, database searching of CE obtains worse result in following query proteins: #6 d1b3ra1, #19 d1d3ga_, #21 d1dbqa_, #22 d1di0a_, #29 d1e4ft1, #32 d1ej8a_, #62 d1i1ra1, #90 d1qfja2, #102 d1ggwa_, #104 d2cmd_1.

There are two reasons to cause the worse result of CE according our observation. First, some retrieval domain proteins have chain-break in their 3D structure files. “Chain-break” means that the residue number is non-continuous in one domain or chain. When the protein occurs this chain-break condition, CE may take this protein as two chains and perform incorrect structure comparison as shown in the Figure 17. Some subject proteins occur this condition in the searching of query proteins, such as #6 d1b3ra1, #21 d1dbqa_, #22 d1di0a_, #104 d2cmd_1. Here, we take subject protein “d1c41a_” in query protein: “#22 d1di0a_” as example, because of the precision of this subject protein in CE is only 0.00813. As shown in Figure 17, there is the condition of chain-break in subject protein “d1c41a_” shown with blue square in Figures 17(A) and (C). The residue number is non-continuous from 76 to 107. The conformation of structure alignment of two proteins is slightly unsatisfied. Furthermore, the alignment length is sorter than the length of query protein and both Z-score and Rmsd is quite low as the alignment result in Figure 17(D). Besides, we observed that CE determines the wrong length of the domain protein “d1c41a_”. The original length of “d1c41a_” is 165 but the size detected by CE is only 72 because of chain-break problem. Nevertheless, PiSA-BLAST is not influenced by chain-break. Even the residue number has been broken; the encoding of structure in PiSA-BLAST method is still continuous.

Second, it is uncertainly that lower Z-score means dissimilar structure. Some protein comparisons possess lower Z-score but present better RMSD. We observed this issue in following query proteins: #19 d1d3ga_, #32 d1ej8a_ and #90 d1qfja2. Here, we take subject protein “d1eso__” in query protein: “#32 d1ej8a_” as example. The precision of this subject protein is only 0.2. In Figure 18, it shows obviously the illustration of the problem of ordering the searching results by Z-score in CE alignment. The comparison of two similar structures is with proper RMSD but displays worse Z-score. It is clearly to see that the comparison between query and subject proteins is not bad. The main secondary structure of these two proteins is aligned appropriately. On the other hand, the gaps inserted into alignment are just loop structure of two proteins. The structures of query and subject proteins are similar and the rmsd is 2.07, but the Z-score is only 4.4. Therefore, the rank of the subject protein is 50 and behind 40 false positive proteins. The performance of CE would be bad in some cases, because we only sorted the retrieval lists by Z-score. We may sort all results that are provided using CE by better way, such as combing Z-score with RMSD.

There is one probably explanation about that PiSA-PSI-BLAST did not enhance supposed performance. Changing the e-value threshold for including sequences in the PSI-BLAST position specific matrix model may cause different alignment results. Although we choose the most appropriate e-value threshold: “10^{-15}”, we may obtain the worse achievement of PiSA-PSI-BLAST in some cases.

For example, there are too many incorrect domain proteins, which are not the same family as query protein, and these e-values of domain proteins are below threshold in searching of query protein “#3 d1ajsa_”. There are 79 subject proteins that are below the e-value threshold. However, there are actually 63 proteins, which are not the same family as query protein. Therefore, the position specific matrix model made by method of PSI-BLAST may include wrong information and cause the iterated searching to go toward wrong result.

On the other hand, there are only a few domain proteins with same family as query below the threshold in several cases. Accordingly, the position specific matrix model may not contain enough sequence information to perform correct searching. For example, there are only 3 proteins below the e-value threshold in searching of query protein “#85 d1pina2”.

PiSA-BLAST can provide the theoretically expected number like e-value of BLAST to indicate what the performance is better. Here, we give 10^{-15} as significance estimate according to our observation. In the Figure 19, the relationship between e-value and structure similarity in PiSA-BLAST is shown. The 1681 points in total on the plot mean every query and subject protein pairs searching in SCOP 95 database. There are 943 points in area (A) and only 79 points in area (B). PiSA-BLAST achieves 98.6% and 92.2% proteins whose Z scores are more than 4.0 and 5.0 when the e-value is less than 10^{-15}.

In Figure 20, it shows the relationship between e-value and precision in PiSA-BLAST. PiSA-BLAST performs 108 queries on the SCOP 95 database. The yellow bars mean that the distribution of e-value of PiSA-BLAST is less than 10^{-15} and red ones mean that the distribution of e-value is more than 10^{-15}. The protein pairs of precision with 80% and upper occupy 91% protein pairs at below 10^{-15} of e-value of PiSA-BLAST. Hence, the value 10^{-15} we given are reasonable.

3.4 Speed Evaluations

Because of the fast sequence alignment method, PiSA-BLAST can search the whole database in a few times. The speed comparison of the selected methods for this experiment is shown in Tables 4 and 5. The results show that PiSA-BLAST can be 5000 times faster than CE in 200 proteins searching. The more amounts of proteins in database, PiSA-BLAST has the more fast speed for searching than CE. It is about 42600 and 34000 times faster than CE in large database and SCOP 95 searching. In addition, the searching times of PiSA-BLAST are near to ones of BLAST. PiSA-BLAST is only about 2.7 times and 4.5 times slower than BLAST respectively in small and large database searching. Another speed comparison of the selected methods for various databases is shown in Table 6. Also, PiSA-BLAST spends execution time about 5 times more than BLAST for whole PDB searching.

3.5 Performance Factor Analysis: Sequence Identity, Structure Similarity and Expect Value

Coded sequence of PiSA-BLAST is not only catch the characteristics of sequence similarity but also hold structure property and information in alignment. Figure 21 shows the percentage of amount of alignment result in each precision rate when sequence identity are equal and more than 25% between query protein and subject ones that are at same SCOP family. On 80% precision and above, PiSA-BLAST do not loss sequence similarity and that is the same as sequence alignment tool: BLAST. Furthermore, PiSA-BLAST possess more accuracy than BLAST in searching SCOP 95% database, a 95% sequence identity filtered subset. Figure 22 shows the percentage of amount of alignment result in each precision rate when sequence identity are less than 25% between query protein and subject ones. In lower sequence identity, BLAST shows worse results in evidence: the percentage of bad precision is partial to high and the one of good precision is low. On the other hand, PiSA-BLAST still holds better accuracy because the percentage of high precision is in the majority.

The relationship between alignment precision and Z-score of the aligned structures similarity is shown in Figures 23 and 24. Structures are superimposed using the CE method and Z-score is calculated. Figures 23 and 24 illustrates that whether structural similarity between query and subject proteins is high or low, PiSA-BLAST presents a more excellent searching precision than BLAST searching. It explains that adding segment information into 1D-sequence can enhance the accuracy of sequence alignment for structure searching.

Figure 25 shows a clear correlation between Z-score (CE) and sequence identity calculated by PiSA-BLAST and BLAST. In Figure 25(A), the correlation between encoded sequence identity of PiSA-BLAST and Z-score of CE displays a better linear relationship. Additionally, the correlation between amino acid sequence identity of BLAST and Z-score shows a worst relationship in Figure 25(B). The correlation coefficient is 0.72 between encoded sequence identity of PiSA-BLAST and Z-score of CE; on the other hand, the correlation coefficient is 0.61 between amino acid sequence identity and Z-score. It is distinct to explain that PiSA-BLAST can take more structural property than BLAST.

3.6 Same Searching Cases Analysis

We give some examples as follows to explain how PiSA-BLAST works and what weakness it has.

Figures 26 to 29 illustrate the results of PiSA-BLAST comparison with 4 SCOP classes of related protein. Here, we took “#89 d1qe0a1”, “#54 d1gr3a_”, “#21 d1dbqa_” and “#16 d1cjwa_" which shown in Table 7 as good examples to searching in SCOP 95 database in order to explain that PiSA-BLAST can provide good performance. We used FASTA [21, 22] program to align two sequences and CE [12] program to perform structural alignment. It is clear to demonstrate that there is quite low sequence identity of original residues between two homologous proteins and high significance estimate of encoded residues after transforming.

Figure 30 is an example to explain one shortcoming in our method for comparison of two related domain proteins. As shown in the Figure 30 (D), the conformation between queries protein "d1mkma1_" in blue and subject one "d1e17a_" in red is similar. However, there are the long structural gaps in protein "d1e17a". These gaps are necessary for structural comparison, but our alignment tool does not allow long gaps to exist. Because there are critical gap open and extension penalties in sequence alignment, the alignment score and e-value of PiSA-BLAST is low in this case.

Figure 31 shows the wrong retrieval of PiSA-BLAST in comparison of two non-related domain proteins. We used PiSA-BLAST to search "d1jbga_" as query. And there is one subject protein, called "d1pk5a_", with high e-value. The SCOP sccs id of "d1jbga_" is a.6.1.3 and SCOP id of "d1pk5a_" is a.123.1.1. Query and subject proteins do not belong to the same family in SCOP, but they have partial structural similarity. The compositions of secondary structure are similar; all of these secondary structures are four to five helices and two short strands. But, the three-dimensional conformation is quite different. Because of consistent composition, the encoded sequence identity and alignment score of PiSA-BLAST is high. The e-value in PiSA-BLAST alignment is $7*10^{-16}$ and the rank of the subject protein is 3.

3.7 PiSA-BLAST on Practical Applications

Figure 32 shows that PiSA-BLAST can be used on the application of fold assignment. There are the result of encoded sequence alignment, structural alignment and 3-dimensional conformation between query protein “1cjw” with A chain and “1wwz” with B chain. The protein “1wwz” is published on 01-Feb-05 and is not assigned in SCOP and CATH currently. PiSA-BLAST is used to assign the fold of the protein “1wwz”, highly similarly to protein “1cjw”, to SCOP sccs id: “d.108.1.1”. The e-value of PiSA-BLAST alignment between “1cjwA” and “1wwzB” is less than 10^{-15}. Then, we performed CE for detail structure alignment and the Z-score of CE alignment is 5.7. Therefore, we suggested that the protein “1wwzB” is assigned the fold of the protein “1cjwA”.

3.8 Web Service

PiSA-BLAST has been setup to a web service as shown in Figure 33. User can input three kinds of query formats: PDB code, SCOP code, and users’ upload 3D structure on the web service to use as a query against the whole structural database and search similar structures. The searching databases includes PDB, nr-PDB, SCOP all, SCOP 95, and SCOP 40. User can acquire the information including the retrieval lists of database searching, the alignment of encoded sequence, the detail structure comparison using CE and the original sequence alignment between query and subject proteins using FASTA [21, 22] program. The hyperlink of our web service is:

http://gemdock.life.nctu.edu.tw/pisa-blast/

Chapter 4

Conclusions

4.1 Summary

In summary, we provide a novel method to do fast one-against-all structural database searching. From 3D to 1D level, PiSA-BLAST can decrease execute time by translating 3D-structure to 1D-sequence and using sequence level to align structure. From 1D to 3D, PiSA-BLAST can enhance the accuracy of sequence alignment for structure searching by adding segment information into 1D-sequence. We use cluster algorithm to group segments, decide representative fragment and assign new codes for structure transforming. After that, we design a rational and usable substitution matrix for new codes. Totally, our results show that PiSA-BLAST is quite efficient and reasonably effective. The database searching time of PiSA-BLAST is very faster then CE. Although PiSA-BLAST could not provide the same accuracy as the results of CE, it can be used as a fast filter to pre-select the top rank 10% to 30% of structure candidates and further evaluation. Given the very fast speed of PiSA-BLAST, this filter-and-refine strategy can reduce the running time by about many folds while maintaining the good accuracy of the detailed comparison methods.

4.2 Major Contributions and Future Perspectives

Here, we have developed a fast structure alignment tool for protein database searching. We evaluated PiSA-BLAST on the retrieval efficiency and effectiveness of the scheme in comparison with the other methods. The results showed that PiSA-BLAST is very much faster than two well-known protein structure comparison methods, DALI and CE and yet not sacrificing on the accuracy of the comparison.

Because PiSA-BLAST can provide a very speedy efficiency on database searching, PiSA-BLAST can be as a useful pre-filtering tool in the near future when the size of protein structure database grows too large to be searched through exhaustively. In filter-and-refine framework, it can be used to reduce the search space before running a more detailed but slower structural comparison method. We are able to perform PiSA-BLAST to do a fast alignment searching at first and output some results of top rank. After that, we achieve the detailed database search by other more delicate but slower structure alignment tools in order to acquiring the best performance and efficiency.

As a future work, we can further improve the accuracy of PiSA-BLAST by using different encoding rules and adding more structural information. Besides, PiSA-BLAST can provide practical applications on fold assignment and homology searching as the preliminary results. Furthermore, our method is to transform 3-dimensional structure to 1D sequence. So, the encoded sequences may be applied to the issue of multiple structure alignment.

Table 1. A small test set selected from previous work [14]. There are 200 members in the database and 20 queries in two SCOP families are listed

Globins family	Serine/ Threonin kinase family
sccs id[a]: a.1.1.2	sccs id: d.144.1.1
d1a6m__	d1a06__
d1ash__	d1apme_
d1b0b__	d1b6cb_
d1fhjb_	d1csn__
d1gcva_	d1f3mc_
d1irda_	d1h8fa_
d1itha_	d1howa_
d1mba__	d1jvpp_
d2gdm__	d1phk__
d3sdha_	d1tkia_

[a] The sccs id is a compact representation of a SCOP domain classification. A sccs identifier includes only the class, fold, superfamily, and family to which each domain belongs to.

Table 2. Summary of 108 queries selected from SCOP all and SCOP 95

SCOP id	SCOP sccs	Query sequence Length	Family Size on 33311 database	Family Size on SCOP 95%	SCOP id	SCOP sccs	Query sequence Length	Family Size on 33311 database	Family Size on SCOP 95%	SCOP id	SCOP sccs	Query sequence Length	Family Size on 33311 database	Family Size on SCOP 95%
d1a8h_2	c.26.1.1	344	48	18	d1euha_	c.82.1.1	470	58	11	d1jz8a4	b.30.1.1	289	124	1
d1afwa2	c.95.1.1	120	114	22	d1exqa_	c.55.3.2	139	61	6	d1k0ia1	c.3.1.2	284	93	13
d1ajsa_	c.67.1.1	408	166	16	d1eyza3	d.142.1.2	195	80	7	d1k94a_	a.39.1.7	161	24	9
d1atg__	c.94.1.1	227	139	26	d1f3mc_	d.144.1.1	279	149	46	d1k9sa_	c.56.2.1	233	86	10
d1aw1a_	c.1.1.1	251	109	15	d1f4pa_	c.23.5.1	143	65	9	d1kbva2	b.6.1.3	147	160	35
d1b3ra1	c.2.1.4	159	48	12	d1f86a_	b.3.4.1	111	100	4	d1kid__	c.8.5.1	189	37	5
d1b5ea_	d.117.1.1	237	154	9	d1fc4a_	c.67.1.4	397	68	18	d1kyga_	c.47.1.10	162	48	16
d1bd3a_	c.61.1.1	220	113	18	d1feca3	d.87.1.1	124	78	18	d1mtyd_	a.25.1.2	508	109	12
d1bg2__	c.37.1.9	319	57	12	d1fjeb2	d.58.7.1	80	63	37	d1kfwa1	c.1.8.5	285	40	17
d1bgva2	c.58.1.1	190	67	9	d1fsoa_	b.1.1.5	134	18	8	d1oela1	a.129.1.1	243	29	3
d1bi5a1	c.95.1.2	231	44	4	d1fxoa_	c.68.1.6	287	70	5	d1onc__	d.5.1.1	100	168	17
d1bu6o1	c.55.1.4	247	46	2	d1g3nc1	a.74.1.1	128	40	12	d1pbga_	c.1.8.4	440	68	11
d1bwvs_	d.73.1.1	134	56	8	d1g5ta_	c.37.1.11	153	121	18	d1pina2	d.26.1.1	115	57	21
d1ccwa_	c.23.6.1	132	41	5	d1g7sa2	b.43.3.1	117	44	11	d1qaxa2	d.179.1.1	306	40	5
d1ce7a_	d.165.1.1	237	58	15	d1geha1	c.1.14.1	294	71	8	d1qdea_	c.37.1.13	193	34	18
d1cjwa_	d.108.1.1	162	42	15	d1ggxa_	d.22.1.1	207	40	6	d1qdlb_	c.23.16.1	191	48	10
d1cp2a_	c.37.1.10	265	107	18	d1gnia3	a.126.1.1	192	80	7	d1qe0a1	c.51.1.1	91	58	9

SCOP id	SCOP sccs	Query sequence Length	Family Size on 33311 database	Family Size on SCOP 95%	SCOP id	SCOP sccs	Query sequence Length	Family Size on 33311 database	Family Size on SCOP 95%	SCOP id	SCOP sccs	Query sequence Length	Family Size on 33311 database	Family Size on SCOP 95%
d1cpt__	a.104.1.1	404	94	14	d1gr3a_	b.22.1.1	128	51	9	d1qfja2	c.25.1.1	131	47	12
d1d3ga_	c.1.4.1	348	53	16	d1gsoa2	c.30.1.1	101	80	6	d1qgna_	c.67.1.3	392	88	15
d1dbfa_	d.79.1.2	123	41	3	d1gtma1	c.2.1.7	235	93	20	d1qgwc_	a.1.1.3	161	66	23
d1dbqa_	c.93.1.1	268	90	16	d1h4vb2	d.104.1.1	292	106	22	d1qkka_	c.23.1.1	135	88	16
d1di0a_	c.16.1.1	144	75	7	d1h8va_	b.29.1.11	214	48	20	d1qmga2	c.2.1.6	222	42	13
d1dk5a_	a.65.1.1	312	50	15	d1hqsa_	c.77.1.1	419	57	9	d1qopb_	c.79.1.1	386	41	9
d1dpga2	d.81.1.5	282	46	3	d1hr6a2	d.185.1.1	233	92	16	d1qora2	c.2.1.1	175	113	17
d1dssg2	d.81.1.1	160	100	16	d1hyha2	d.162.1.1	150	97	24	d1qq4a_	b.47.1.1	194	64	12
d1dzka_	b.60.1.1	144	107	26	d1i1ra1	b.1.2.1	96	136	55	d1smva_	b.10.1.2	192	15	13
d1e0ta2	c.1.12.1	215	66	5	d1idsa2	d.44.1.1	110	106	17	d1trb_2	c.3.1.5	121	177	46
d1e0ta3	c.49.1.1	113	65	5	d1ie9a_	a.123.1.1	251	96	29	d1vcaa2	b.1.1.4	86	159	61
d1e4ft1	c.55.1.1	189	91	23	d1ig8a1	c.55.1.3	202	40	12	d1vdra_	c.71.1.1	153	116	10
d1e6ta_	d.85.1.1	125	121	5	d1ih7a1	c.55.3.5	371	47	13	d1ggwa_	a.39.1.5	138	127	41
d1eal__	b.60.1.2	123	63	25	d1is8a_	d.96.1.1	184	90	2	d1zin_1	c.37.1.1	174	139	31
d1ej8a_	b.1.8.1	136	83	12	d1j7na3	d.166.1.1	257	50	9	d2cmd_1	c.2.1.5	141	97	26
d1ekxa1	c.78.1.1	146	179	15	d1jb7a2	b.40.4.3	120	41	22	d2shpa1	c.45.1.2	263	49	12
d1ep3b1	b.43.4.2	97	50	17	d1jjwa_	d.153.1.4	169	164	35	d1cqda_	d.3.1.1	454	92	28
d1eu3a1	b.40.2.2	93	89	14	d1hrha1	c.55.3.1	148	93	11	d3grx__	c.47.1.1	77	63	19
d1euaa_	c.1.10.1	209	130	20	d1jswa_	a.127.1.1	455	45	11	d3pmga1	c.84.1.1	186	54	9

Table 3. Comparison PiSA-BLAST with six methods on the dataset shown in Table 1

No. of relevant	Average no. of retrievals required[b]						
retrievals[a]	DALI[c]	CE[c]	TopScan[c]	ProtDex2[c]	BLAST	PSI-BLAST	PiSA-BLAST
1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
4	4	4	5	4	7	7	4
6	6	6	8	6	18	17	6
8	8	8	14	9	47	25	8
10	10	10	29	16	93	38	10

[a] Relevant retrieval is defined as an event of retrieving a protein from the database that belongs to the same 'family' as the query.

[b] The number represents the average ranking under the various methods to retrieve the number of relevant answers in [a].

[c] The results are directly summarized from [14] .

Table 4. Executing times of 20 queries on the database with 200 proteins shown in Table 1

Method	Total time [a] (in seconds)	Average time per query (in seconds)	Related ratio comparing to PiSA-BLAST [b]
DALI [c]	23464	1173.180	25014
CE [c]	4632	231.600	4938
TopScan [c]	7.310	0.366	7.79
ProtDex2 [c]	1.982	0.099	2.11
BLAST	0.335	0.0168	0.36
PSI-BLAST	1.052	0.0526	1.12
PiSA-BLAST	0.938	0.0469	1.00

[a] The total searching time of every query searching in the small database.

[b] The ratio of total time of PiSA-BLAST to various methods.

[c] The results are directly summarized from [14].

Table 5. Running times of 108 queries on the database with 33311 proteins shown in Table 2

Method	Total time[a] (in seconds)	Average time per query (in seconds)	Related ratio comparing to PiSA-BLAST[b]
DALI[c]	about 250 days	about 2.31 days	about 216000
CE[c]	about 50 days	about 0.46 days	about 43000
TopScan[d]	11715	108.475	117
ProtDex2[d]	104	0.967	1.05
BLAST	22.196	0.2055	0.22
PSI-BLAST	53.722	0.4974	0.54
PiSA-BLAST	99.901	0.9250	1.00

[a] The total searching time of every query searching in the large database with 33311 proteins.

[b] The ratio of total time of PiSA-BLAST to various methods.

[c] The total searching time of DALI and CE is approximate time.

[d] The results are directly summarized from [14].

Table 6. Comparison running times of BLAST, PSI-BLAST and PiSA-BLAST for 108 queries searching on five databases selected from PDB and SCOP. These 108 queries are shown in Table 2

Database	Published date	Number of sequence in database	Total running times (in seconds)		
			BLAST	PSI-BLAST	PiSA-BLAST
PDB	19-Apr-05	64333	53.517	119.444	240.774
nr-PDB	19-Apr-05	10308	9.164	23.883	35.050
SCOP	1.65	53659	34.452	76.092	155.178
SCOP 95%	1.65	9354	6.921	18.312	34.349
SCOP 40%	1.65	5630	4.713	13.163	22.487

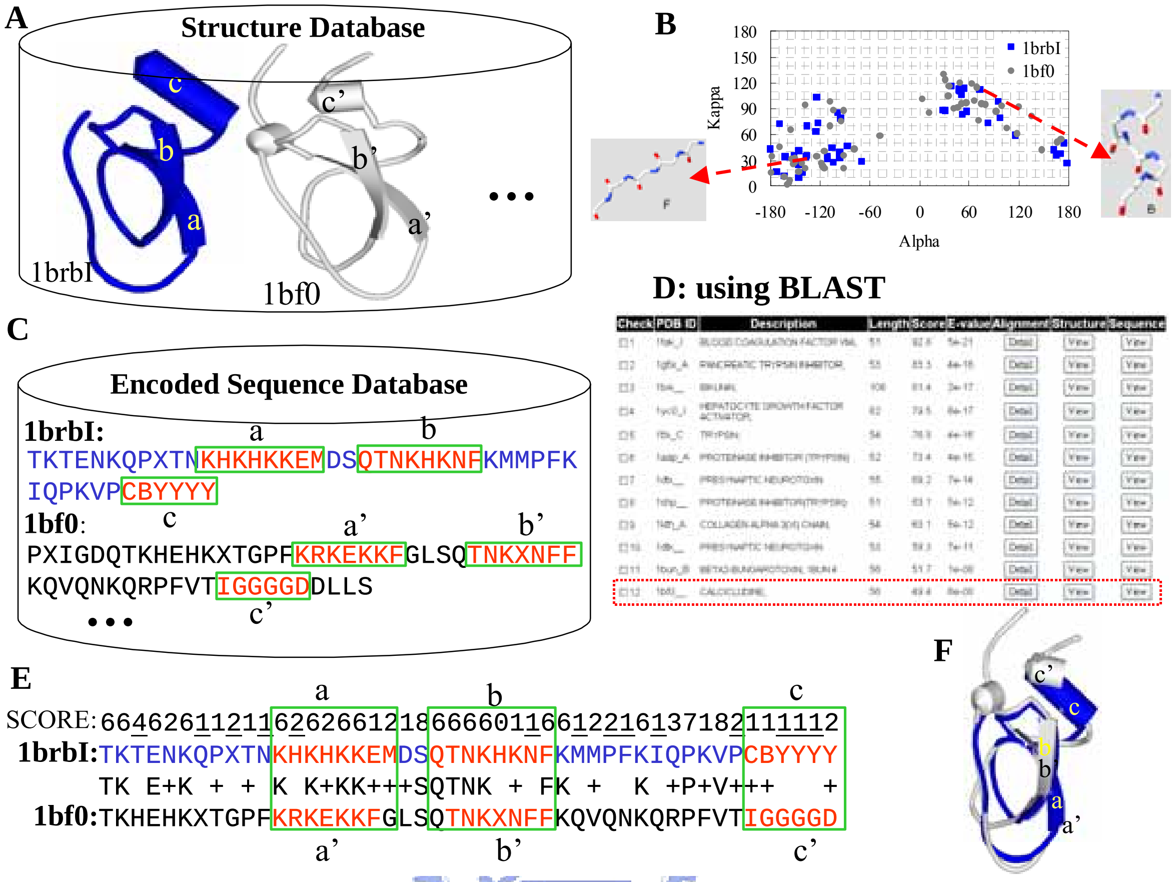
Table 7. Average precisions of five alignment tools on 108 queries searching on the SCOP 95 database. These 108 queries are shown in Table 2

Query #	SCOP id	SCOP sccs	One-code class ID	Query sequence Length	Family Size on SCOP 95%	Average precision				
						CE	BLAST	PSI-BLAST	PiSA-BLAST	PiSA-PSI-BLAST
1	d1a8h_2	c.26.1.1	C	344	18	0.905	0.470	0.802	0.708	0.707
2	d1afwa2	c.95.1.1	C	120	22	0.734	0.138	0.325	0.521	0.666
3	d1ajsa_	c.67.1.1	C	408	16	0.886	0.591	0.897	**0.901**	0.884
4	d1atg__	c.94.1.1	C	227	26	0.888	0.125	0.213	0.717	0.718
5	d1awla_	c.1.1.1	C	251	15	1.000	1.000	1.000	1.000	1.000
6	d1b3ra1	c.2.1.4	C	159	12	0.556	0.279	0.828	**0.746**	0.746
7	d1b5ea_	d.117.1.1	D	237	9	1.000	0.595	1.000	1.000	1.000
8	d1bd3a_	c.61.1.1	C	220	18	0.734	0.200	0.355	0.632	0.633
9	d1bg2__	c.37.1.9	C	319	12	0.507	0.511	0.530	**0.579**	0.581
10	d1bgva2	c.58.1.1	C	190	9	1.000	0.858	1.000	1.000	1.000
11	d1bi5a1	c.95.1.2	C	231	4	0.530	0.504	0.501	0.502	0.502
12	d1bu6o1	c.55.1.4	C	247	2	0.500	0.501	0.501	0.500	0.500
13	d1bwvs_	d.73.1.1	D	134	8	1.000	1.000	1.000	1.000	1.000
14	d1ccwa_	c.23.6.1	C	132	5	1.000	0.800	0.800	0.819	0.820
15	d1ce7a_	d.165.1.1	D	237	15	1.000	1.000	1.000	1.000	1.000
16	d1cjwa_	d.108.1.1	D	162	15	0.967	0.149	0.348	**0.972**	0.972
17	d1cp2a_	c.37.1.10	C	265	18	0.561	0.405	0.582	0.538	0.539
18	d1cpt__	a.104.1.1	A	404	14	1.000	1.000	1.000	1.000	1.000
19	d1d3ga_	c.1.4.1	C	348	16	0.547	0.410	0.512	**0.650**	0.653
20	d1dbfa_	d.79.1.2	D	123	3	1.000	1.000	1.000	1.000	1.000
21	d1dbqa_	c.93.1.1	C	268	16	0.840	0.391	0.648	0.952	0.947
22	d1di0a_	c.16.1.1	C	144	7	0.858	1.000	1.000	1.000	1.000
23	d1dk5a_	a.65.1.1	A	312	15	1.000	0.933	1.000	0.935	0.935
24	d1dpga2	d.81.1.5	D	282	3	1.000	0.667	0.667	0.667	0.667
25	d1dssg2	d.81.1.1	D	160	16	1.000	0.690	0.730	0.924	0.927
26	d1dzka_	b.60.1.1	B	144	26	0.986	0.410	0.789	0.899	0.899
27	d1e0ta2	c.1.12.1	C	215	5	1.000	1.000	1.000	1.000	1.000
28	d1e0ta3	c.49.1.1	C	113	5	1.000	1.000	1.000	1.000	1.000
29	d1e4ft1	c.55.1.1	C	189	23	0.057	0.056	0.058	0.089	0.089
30	d1e6ta_	d.85.1.1	D	125	5	1.000	0.646	0.702	1.000	1.000
31	d1eal__	b.60.1.2	B	123	25	1.000	0.945	1.000	0.966	0.975
32	d1ej8a_	b.1.8.1	B	136	12	0.501	0.423	0.424	**0.994**	1.000
33	d1ekxa1	c.78.1.1	C	146	15	0.471	0.470	0.468	0.470	0.470

Query #	SCOP id	SCOP sccs	One-code class ID	Query sequence Length	Family Size on SCOP 95%	Average precision				
						CE	BLAST	PSI-BLAST	PiSA-BLAST	PiSA-PSI-BLAST
34	d1ep3b1	b.43.4.2	B	97	17	0.965	0.078	0.078	0.774	0.772
35	d1eu3a1	b.40.2.2	B	93	14	0.743	0.220	0.520	0.677	0.677
36	d1euaa_	c.1.10.1	C	209	20	0.184	0.058	0.057	0.165	0.165
37	d1euha_	c.82.1.1	C	470	11	1.000	1.000	1.000	1.000	1.000
38	d1exqa_	c.55.3.2	C	139	6	1.000	1.000	1.000	1.000	1.000
39	d1eyza3	d.142.1.2	D	195	7	0.957	0.584	0.779	0.857	0.857
40	d1f3mc_	d.144.1.1	D	279	46	1.000	1.000	1.000	1.000	1.000
41	d1f4pa_	c.23.5.1	C	143	9	0.836	0.855	0.854	0.772	0.772
42	d1f86a_	b.3.4.1	B	111	4	1.000	1.000	1.000	1.000	1.000
43	d1fc4a_	c.67.1.4	C	397	18	0.582	0.312	0.608	0.569	0.566
44	d1feca3	d.87.1.1	D	124	18	0.899	0.775	0.834	0.853	0.853
45	d1fjeb2	d.58.7.1	D	80	37	0.928	0.732	0.999	0.824	0.826
46	d1fsoa_	b.1.1.5	B	134	8	0.986	0.877	0.875	0.986	0.986
47	d1fxoa_	c.68.1.6	C	287	5	1.000	1.000	1.000	1.000	1.000
48	d1g3nc1	a.74.1.1	A	128	12	0.744	0.352	0.336	0.429	0.429
49	d1g5ta_	c.37.1.11	C	153	18	0.457	0.130	0.130	0.341	0.341
50	d1g7sa2	b.43.3.1	B	117	11	0.940	0.092	0.092	0.666	**0.956**
51	d1geha1	c.1.14.1	C	294	8	1.000	1.000	1.000	1.000	1.000
52	d1ggxa_	d.22.1.1	D	207	6	1.000	1.000	1.000	1.000	1.000
53	d1gnia3	a.126.1.1	A	192	7	0.957	0.866	1.000	0.860	0.860
54	d1gr3a_	b.22.1.1	B	128	9	0.895	0.224	0.224	0.890	**1.000**
55	d1gsoa2	c.30.1.1	C	101	6	0.611	0.169	0.169	0.438	0.438
56	d1gtma1	c.2.1.7	C	235	20	0.860	0.406	0.469	0.733	0.744
57	d1h4vb2	d.104.1.1	D	292	22	0.569	0.162	0.411	0.634	0.647
58	d1h8va_	b.29.1.11	B	214	20	1.000	0.529	0.446	0.986	0.986
59	d1hqsa_	c.77.1.1	C	419	9	1.000	1.000	1.000	1.000	1.000
60	d1hr6a2	d.185.1.1	D	233	16	0.979	0.447	0.442	0.871	0.872
61	d1hyha2	d.162.1.1	D	150	24	1.000	0.791	1.000	1.000	1.000
62	d1i1ra1	b.1.2.1	B	96	55	0.055	0.024	0.025	0.028	0.028
63	d1idsa2	d.44.1.1	D	110	17	1.000	1.000	1.000	1.000	1.000
64	d1ie9a_	a.123.1.1	A	251	29	1.000	1.000	1.000	1.000	1.000
65	d1ig8a1	c.55.1.3	C	202	12	0.642	0.501	0.501	0.505	0.505
66	d1ih7a1	c.55.3.5	C	371	13	0.880	0.463	0.464	0.549	0.532
67	d1is8a_	d.96.1.1	D	184	2	1.000	1.000	1.000	1.000	1.000
68	d1j7na3	d.166.1.1	D	257	9	0.667	0.125	0.124	0.667	0.668

Query #	SCOP id	SCOP sccs	One-code class ID	Query sequence Length	Family Size on SCOP 95%	Average precision				
						CE	BLAST	PSI-BLAST	PiSA-BLAST	PiSA-PSI-BLAST
69	d1jb7a2	b.40.4.3	B	120	22	0.502	0.099	0.100	0.301	0.302
70	d1jjwa_	d.153.1.4	D	169	35	1.000	0.350	0.684	0.999	0.999
71	d1hrha1	c.55.3.1	C	148	11	0.729	0.628	0.661	0.793	0.796
72	d1jswa_	a.127.1.1	A	455	11	1.000	0.965	1.000	1.000	1.000
73	d1jz8a4	b.30.1.1	B	289	1	1.000	1.000	1.000	1.000	1.000
74	d1k0ia1	c.3.1.2	C	284	13	0.343	0.325	0.320	0.206	0.229
75	d1k94a_	a.39.1.7	A	161	9	1.000	1.000	1.000	1.000	1.000
76	d1k9sa_	c.56.2.1	C	233	10	1.000	0.504	0.719	1.000	1.000
77	d1kbva2	b.6.1.3	B	147	35	0.893	0.144	0.179	0.832	0.816
78	d1kid__	c.8.5.1	C	189	5	1.000	1.000	1.000	1.000	1.000
79	d1kyga_	c.47.1.10	C	162	16	1.000	0.457	0.700	0.740	0.739
80	d1mtyd_	a.25.1.2	A	508	12	0.918	0.172	0.172	0.633	0.637
81	d1kfwa1	c.1.8.5	C	285	17	0.911	0.662	0.791	0.867	0.867
82	d1oela1	a.129.1.1	A	243	3	1.000	1.000	1.000	1.000	1.000
83	d1onc__	d.5.1.1	D	100	17	1.000	0.975	1.000	1.000	1.000
84	d1pbga_	c.1.8.4	C	440	11	1.000	1.000	1.000	1.000	1.000
85	d1pina2	d.26.1.1	D	115	21	0.341	0.292	0.300	0.219	0.219
86	d1qaxa2	d.179.1.1	D	306	5	1.000	1.000	1.000	1.000	1.000
87	d1qdea_	c.37.1.13	C	193	18	0.486	0.209	0.229	0.301	0.302
88	d1qdlb_	c.23.16.1	C	191	10	1.000	0.878	1.000	1.000	1.000
89	d1qe0a1	c.51.1.1	C	91	9	1.000	0.556	0.927	1.000	1.000
90	d1qfja2	c.25.1.1	C	131	12	**0.610**	0.252	**0.835**	0.746	0.746
91	d1qgna_	c.67.1.3	C	392	15	0.808	0.568	0.696	0.741	0.735
92	d1qgwc_	a.1.1.3	A	161	23	1.000	1.000	1.000	1.000	1.000
93	d1qkka_	c.23.1.1	C	135	16	0.913	0.875	**0.992**	0.808	0.809
94	d1qmga2	c.2.1.6	C	222	13	0.715	0.177	0.236	0.266	0.124
95	d1qopb_	c.79.1.1	C	386	9	1.000	0.766	0.931	0.932	0.932
96	d1qora2	c.2.1.1	C	175	17	1.000	0.837	1.000	1.000	1.000
97	d1qq4a_	b.47.1.1	B	194	12	0.671	0.537	0.634	**0.704**	0.685
98	d1smva_	b.10.1.2	B	192	13	1.000	1.000	1.000	1.000	1.000
99	d1trb_2	c.3.1.5	C	121	46	0.623	0.257	0.613	0.541	0.542
100	d1vcaa2	b.1.1.4	B	86	61	0.233	0.113	0.229	0.173	0.086
101	d1vdra_	c.71.1.1	C	153	10	1.000	1.000	1.000	1.000	1.000
102	d1ggwa_	a.39.1.5	A	138	41	0.468	0.803	0.833	0.743	0.746
103	d1zin_1	c.37.1.1	C	174	31	0.601	0.369	0.694	0.622	0.623

Query #	SCOP id	SCOP sccs	One-code class ID	Query sequence Length	Family Size on SCOP 95%	Average precision				
						CE	BLAST	PSI-BLAST	PiSA-BLAST	PiSA-PSI-BLAST
104	d2cmd_1	c.2.1.5	C	141	26	0.834	0.887	0.992	**0.995**	0.995
105	d2shpa1	c.45.1.2	C	263	12	1.000	0.988	1.000	1.000	1.000
106	d1cqda_	d.3.1.1	D	454	28	0.974	0.929	0.982	0.960	0.960
107	d3grx__	c.47.1.1	C	77	19	0.347	0.327	0.486	0.267	0.268
108	d3pmga1	c.84.1.1	C	186	9	0.412	0.335	0.334	0.337	0.337


A
Structure Database
c
b
a
1brbI
c'
b'
a'
1bf0
•••
B
180
150
120
90
60
30
0
Kappa
-180
-120
-60
0
60
120
180
Alpha
1brbI
1bf0
F
B
C
Encoded Sequence Database
1brbI: a b
TKTENKQPXTNKHKHKKEMDSQTNKHKNFKMMPFK
IQPKVPCBYYYY
1bf0: c a' b'
PXIGDQTKHEHKXTGPFKRKEKKFGLSQTNKXNFF
KQVQNKQRPFVTIGGGGDDLLS
••• c'
D: using BLAST
E
a b c
SCORE: 66462611211626266121866660116612216137182111112
1brbI:TKTENKQPXTNKHKHKKEMDSQTNKHKNFKMMPFKIQPKVPCBYYYY
TK E+K + + K K+KK+++SQTNK + FK + K +P+V+++ +
1bf0:TKHEHKXTGPFKRKEKKFGLSQTNKXNFFKQVQNKQRPFVTIGGGGD
a' b' c'
F
c'
c
b
b'
a
a'

Check	PDB ID	Description	Length	Score	E-value	Alignment	Structure	Sequence
□1	1fak_I	BLOOD COAGULATION FACTOR VIIA	51	92.8	5e-21	Detail	View	View
□2	1gfx_A	PANCREATIC TRYPSIN INHIBITOR	53	83.3	4e-18	Detail	View	View
□3	1bik__	BIKUNIN	108	81.4	3e-17	Detail	View	View
□4	1yc0_I	HEPATOCYTE GROWTH FACTOR ACTIVATOR	62	79.5	8e-17	Detail	View	View
□5	1fjs_C	TRYPSIN	54	76.9	4e-16	Detail	View	View
□6	1aap_A	PROTEINASE INHIBITOR (TRYPSIN)	52	73.4	4e-15	Detail	View	View
□7	1dtx__	PRESYNAPTIC NEUROTOXIN	58	69.2	7e-14	Detail	View	View
□8	1shp__	PROTEINASE INHIBITOR(TRYPSIN)	51	63.1	5e-12	Detail	View	View
□9	1kth_A	COLLAGEN ALPHA 3(VI) CHAIN	54	63.1	5e-12	Detail	View	View
□10	1dtk__	PRESYNAPTIC NEUROTOXIN	53	59.3	7e-11	Detail	View	View
□11	1bun_B	BETA2-BUNGAROTOXIN, SBUN 4	58	51.7	1e-08	Detail	View	View
□12	1bf0__	CALCICLUDINE	58	49.4	8e-08	Detail	View	View

Figure 1. Step-by-step illustration of the PiSA-BLAST methodology using 1brbI as the query protein searching against nr-PDB (protein data bank). (A) The two structures (1brbI is blue and 1bf0 is gray) to be compared showing protein structures. (**B**) The kappa-alpha angle (κ, α) 2D map of all residues in each of the two proteins. These two proteins have the similar (κ, α) 2D maps. (**C**) All of 3D-protein structures in the nr-PDB are encoded into 1D-structure sequences with 23 different codes according to the (κ, α) 2D map (see text). The red codes are the SSE parts in each of the two proteins. (**D**) The structure searching results using BLAST with our new substitution matrix (see text). (**E**) The aligned result and score of two1D-structure sequences. The score is calculated according to the substitution matrix, e.g., the score is 6 aligning 'T' to 'T', 6 aligning 'K' to 'K', and –4 aligned 'T' to 'H'. (**F**) The resulting structure alignments for the alignment solution identified in (E) by structure alignment tool, CE.

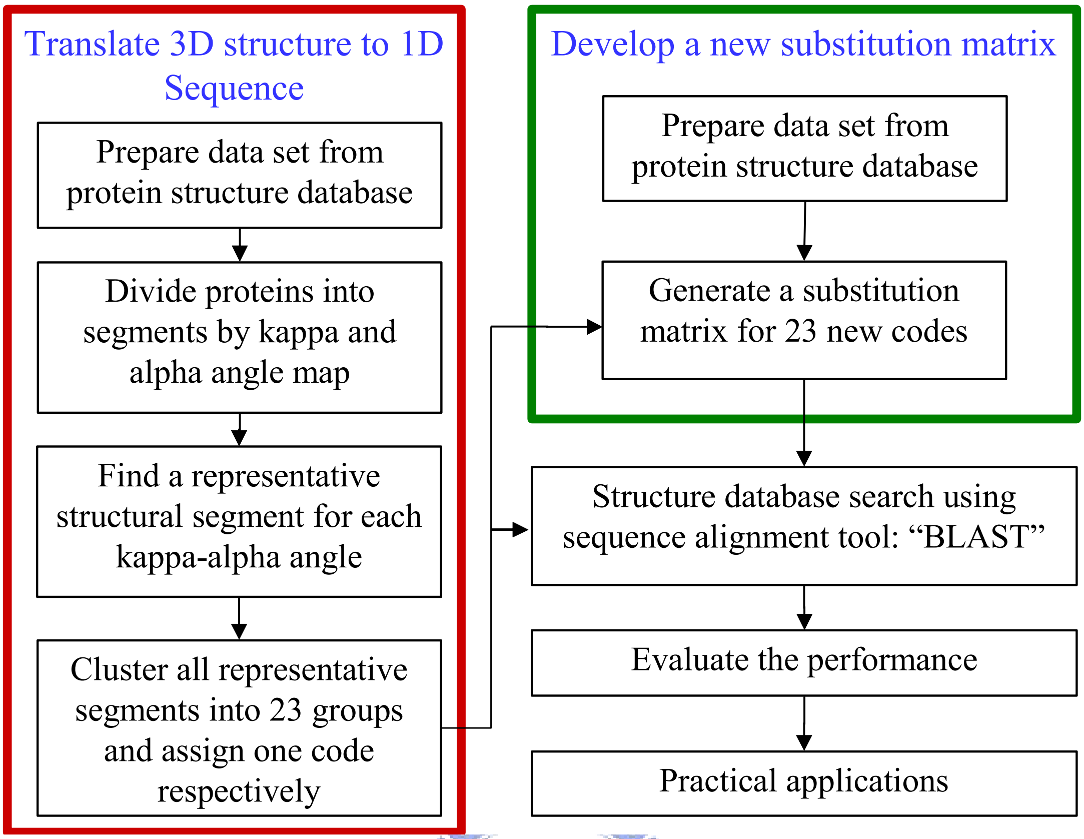

Translate 3D structure to 1D Sequence
Prepare data set from protein structure database
Divide proteins into segments by kappa and alpha angle map
Find a representative structural segment for each kappa-alpha angle
Cluster all representative segments into 23 groups and assign one code respectively
Develop a new substitution matrix
Prepare data set from protein structure database
Generate a substitution matrix for 23 new codes
Structure database search using sequence alignment tool: "BLAST"
Evaluate the performance
Practical applications

Figure 2. Overview of our method. First, we prepare training set from ASTRAL SCOP database 1.65 40% set. Second, we divide domain proteins of training set into many segments that are have various kappa and alpha angle. Then, we find representative segments of each kappa and alpha angle and use cluster algorithm to group these representative segments. After that, we assign a new code for each representative group. Next, we need to develop a substitution matrix for new codes and use it to replace default matrix for sequence alignment tool. We can use sequence alignment tool to do fast protein structure searching in database and evaluate the performance. Finally, we apply the PiSA-BLAST on practical application.

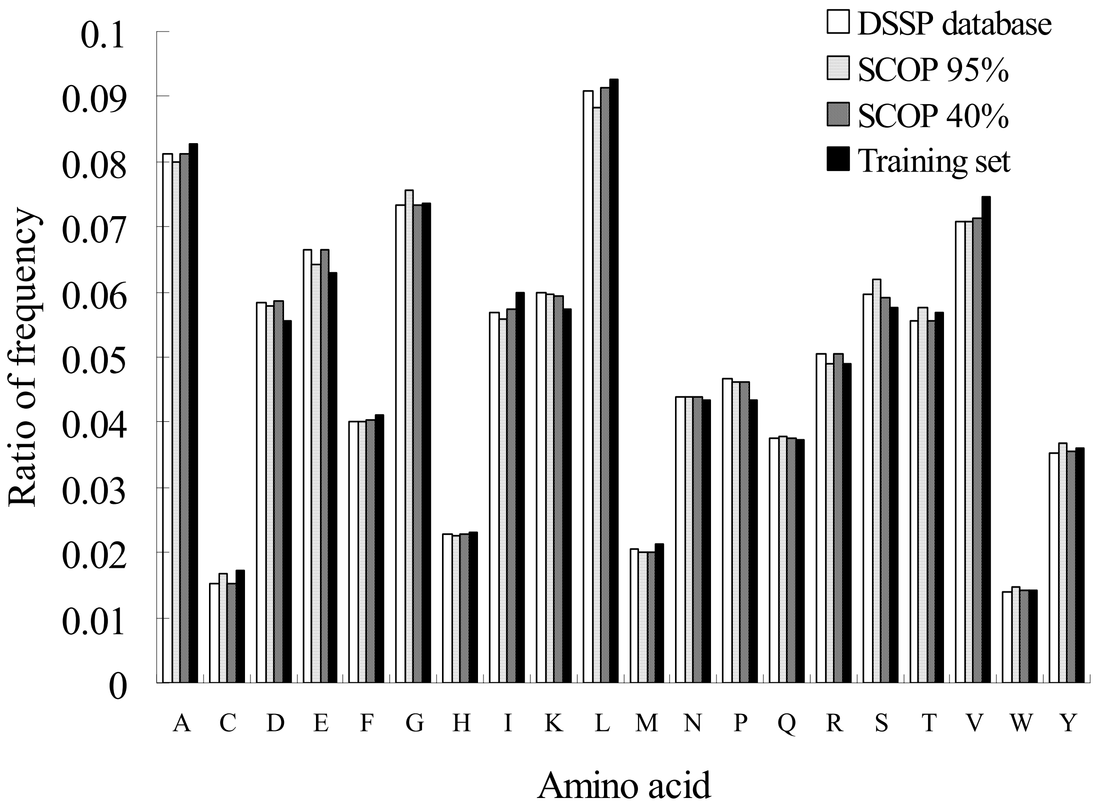


Figure 3. Comparison the amino acids compositions of our train set, including 1584 proteins for encoding the structured codes and the substitute matrix, with three well-known structure databases (DSSP database, SCOP 95 and SCOP 40 database). The distributions of amino acids compositions of these four databases are similar.

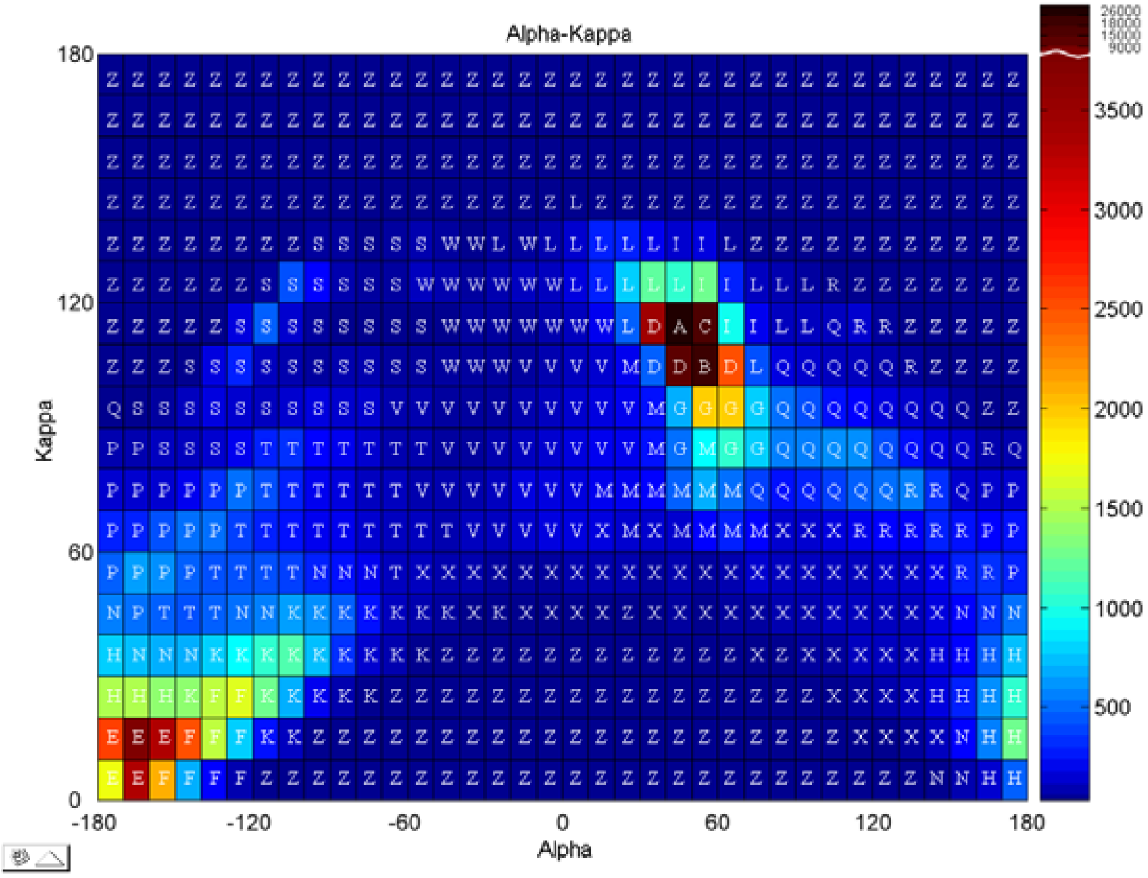


Figure 4. The kappa-alpha distribution of 263696 segments in our training set (792 protein pairs) are colored. The color bar on the right side shows the distribution scale. These segments are encoded into 23 codes based on the distributions of kappa and alpha angle. The helix-like segments (e.g., A, B, C and D) have more than 9000 segments whose alpha angle ranging from 40° to 60° and kappa angle ranging from 100° to 120°. The strand-like segments (e.g., E and F) have over 3000 segments with alpha angle ranging from -180° to -140° and kappa angle ranging from 0° to 20°.

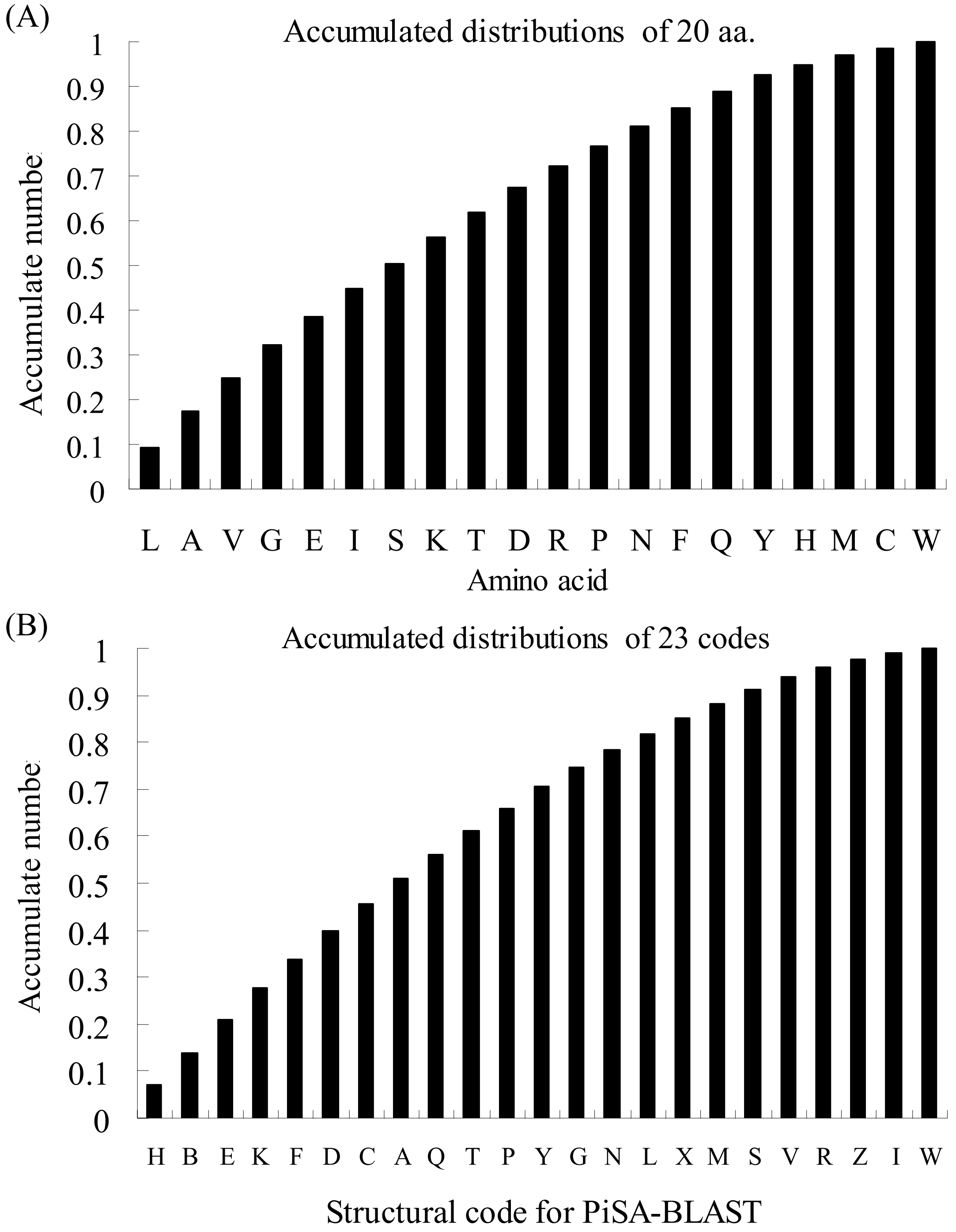


Figure 5. Accumulated distributions of (A) 20 kinds of amino acids and (B) 23 new codes in training set. The accumulated distribution of 23 codes is similar to the distribution of 20 amino acids. The most number in 20 amino acids is amino acid, leucine (L), and the ratio is 9.26%. The most quantity in 23 new codes for PiSA-BLAST is H and the ratio is 6.99%.

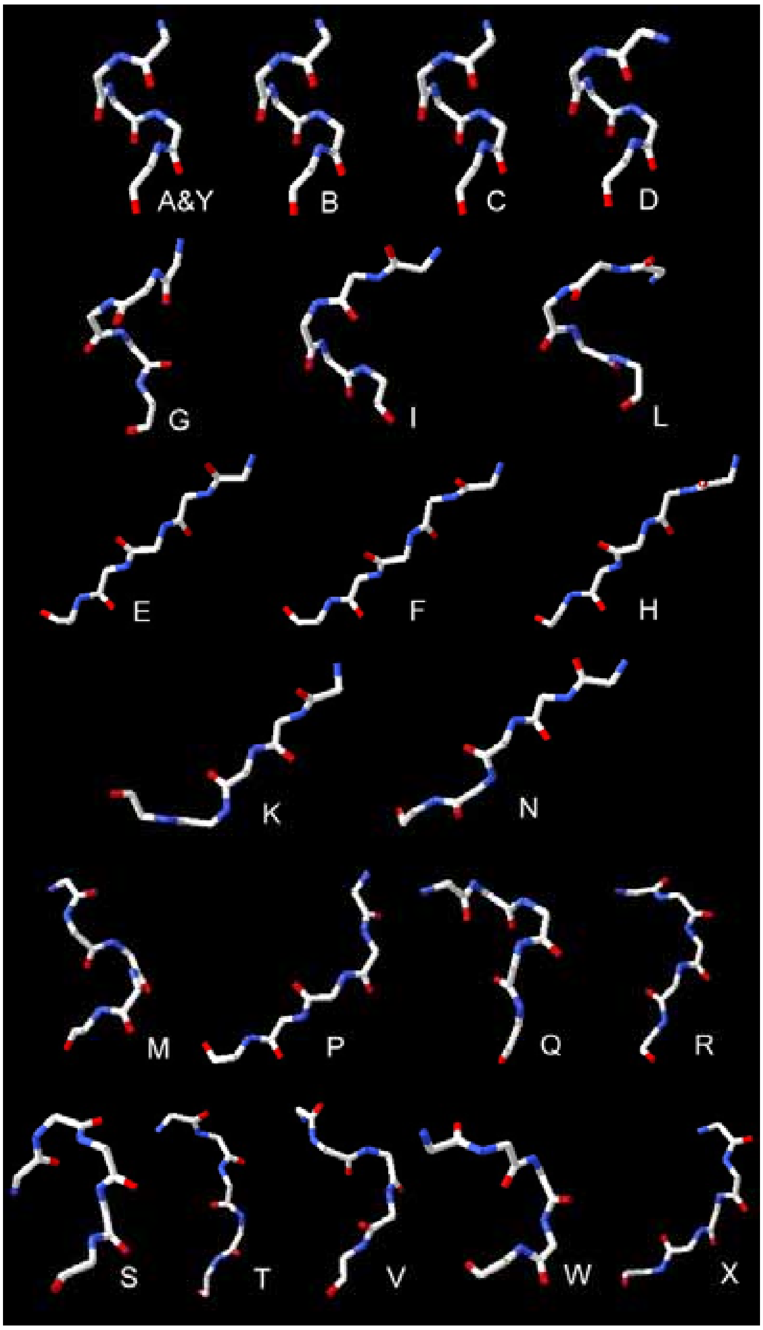


Figure 6. The conformations of the representative segments of 23 new codes. The new codes, A, Y, B, C and D, are helix; G, I and L are helix-like; F and H are strand; K and N are strand-like; and the other codes are loop-like segments.

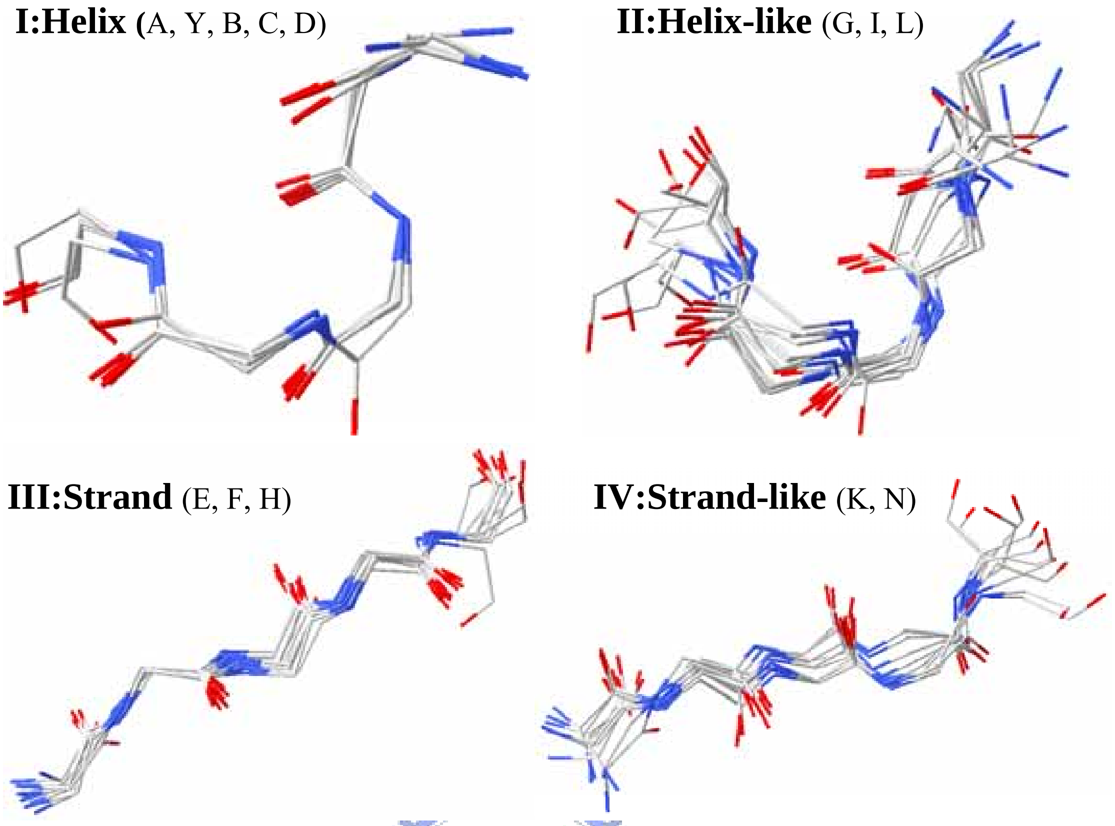
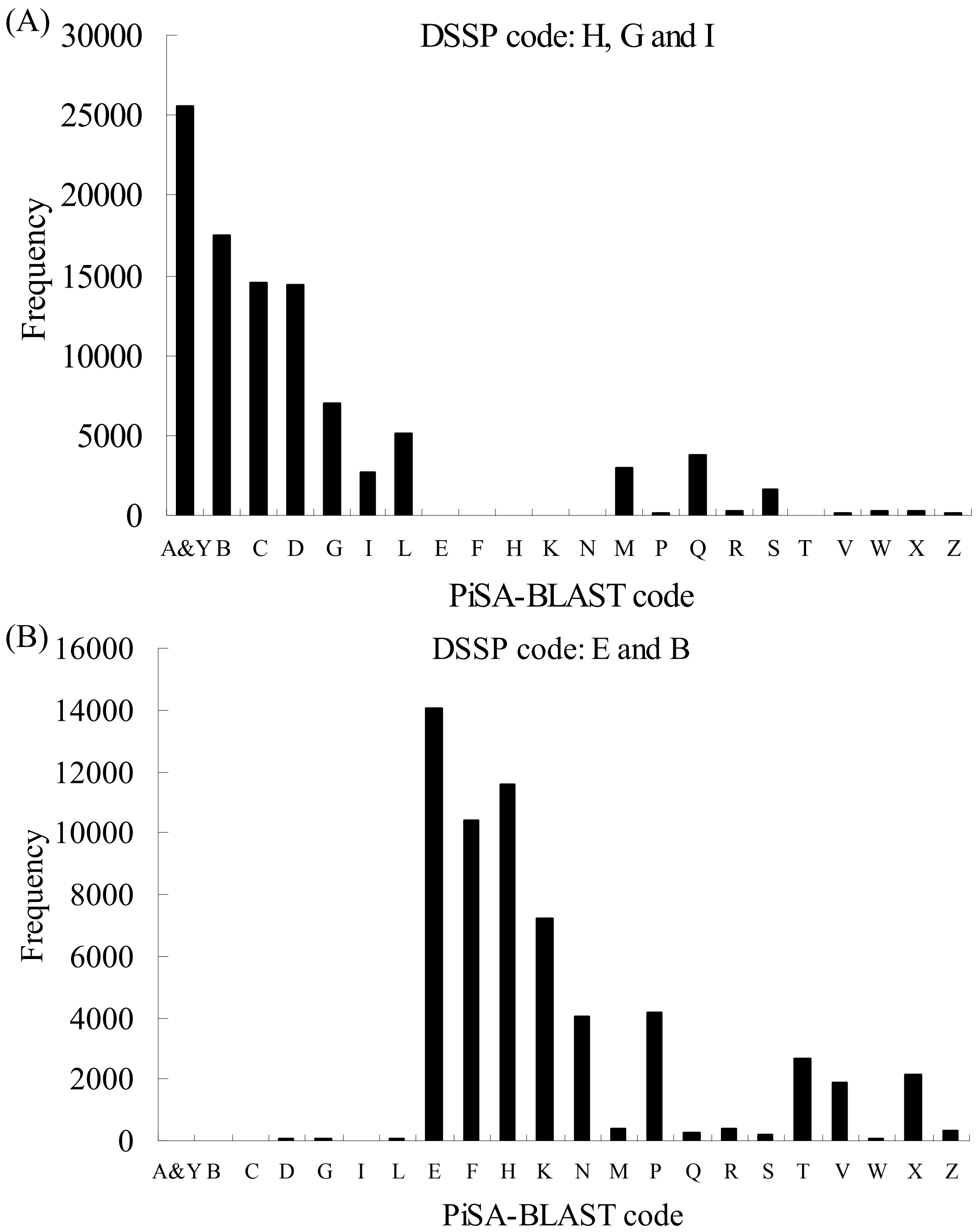


Figure 7. The conformations of representative segment in each cell of four main groups: (I) helix codes (A, Y, B, C, D) have 4 segments; (II) helix-like codes (G, I, L) have 12 segments; (III) strand codes (E, F, H) have 15 segments; (IV) strand-like codes (K, N) have 11 segments. As the conformations show, the structure of segments is very similar in same secondary structure defined region.


(A)
DSSP code: H, G and I
Frequency
30000
25000
20000
15000
10000
5000
0
A&Y B C D G I L E F H K N M P Q R S T V W X Z
PiSA-BLAST code
(B)
DSSP code: E and B
Frequency
16000
14000
12000
10000
8000
6000
4000
2000
0
A&Y B C D G I L E F H K N M P Q R S T V W X Z
PiSA-BLAST code

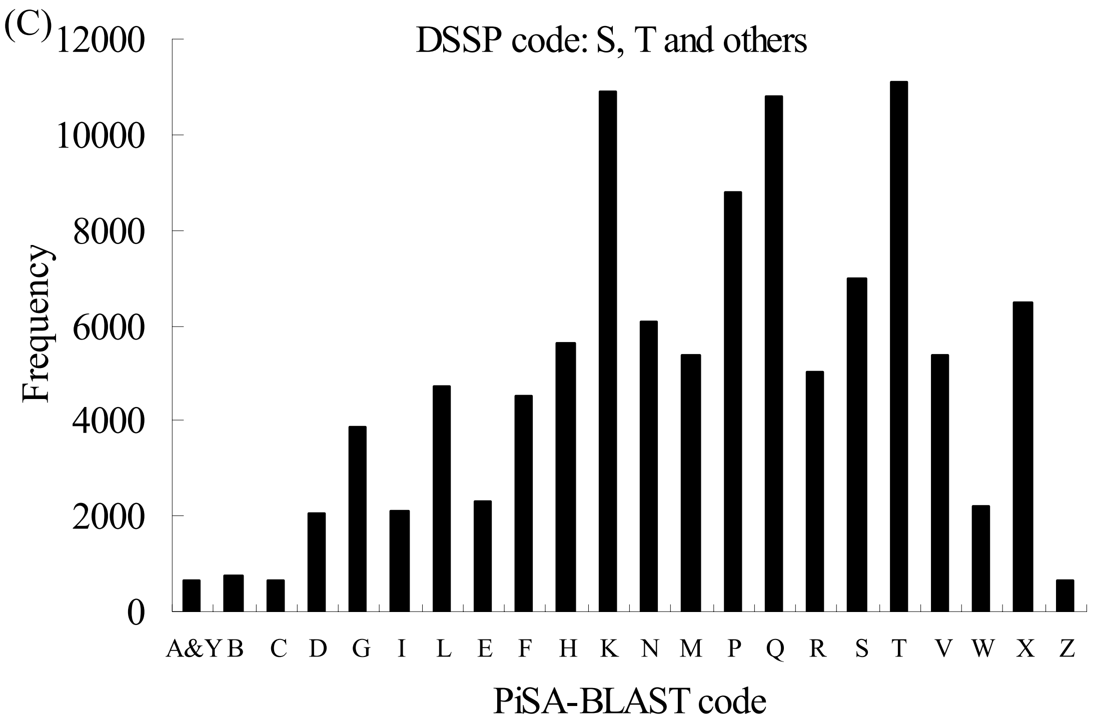


Figure 8. The distribution relationship between 23 new codes (in PiSA-BLAST) and 8 secondary structure codes (in DSSP): (A) The structural-coded distribution of helix codes (H, G and I) in DSSP; (B) The structural-coded distribution of strand codes (E and B) in DSSP; (C) The structural-coded distribution of loop codes (S, T and others) in DSSP. The distributions of helix, helix-like, strand and strand-like segments defined by PiSA-BLAST are high related to secondary structures in DSSP.

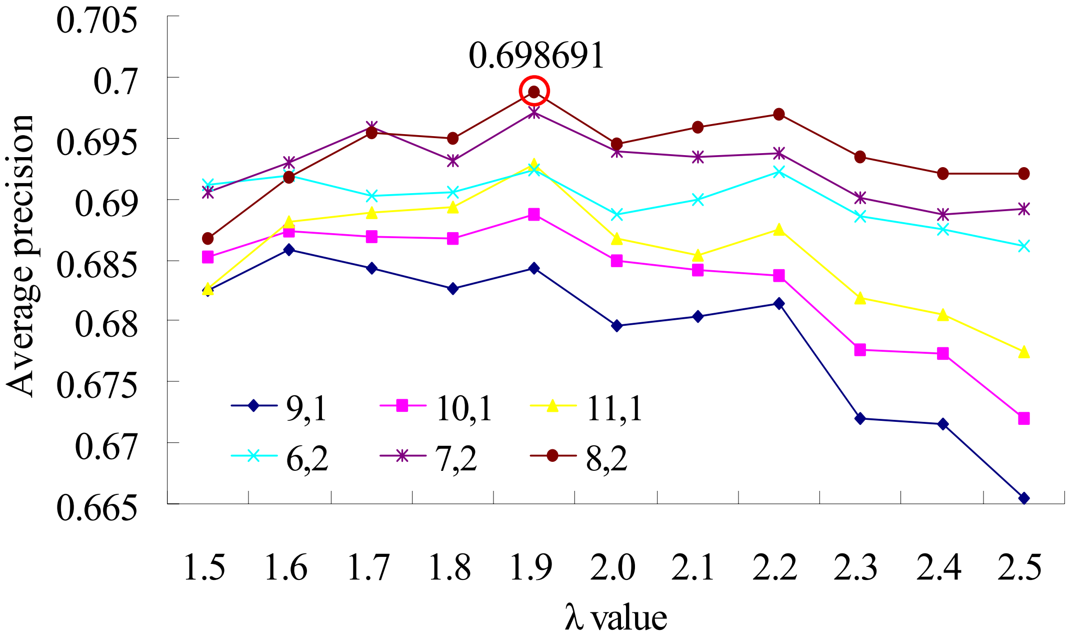


Figure 9. The average precisions of PiSA-BLAST on 108 queries searching on SCOP 95 using various values of λ and gap penalty. We tested six kinds values of open and extend gap penalty with different λ values to find out the optimized parameter for the performance of PiSA-BLAST. Here, the open gap penalty is set to 8 and extend gap penalty is 2.

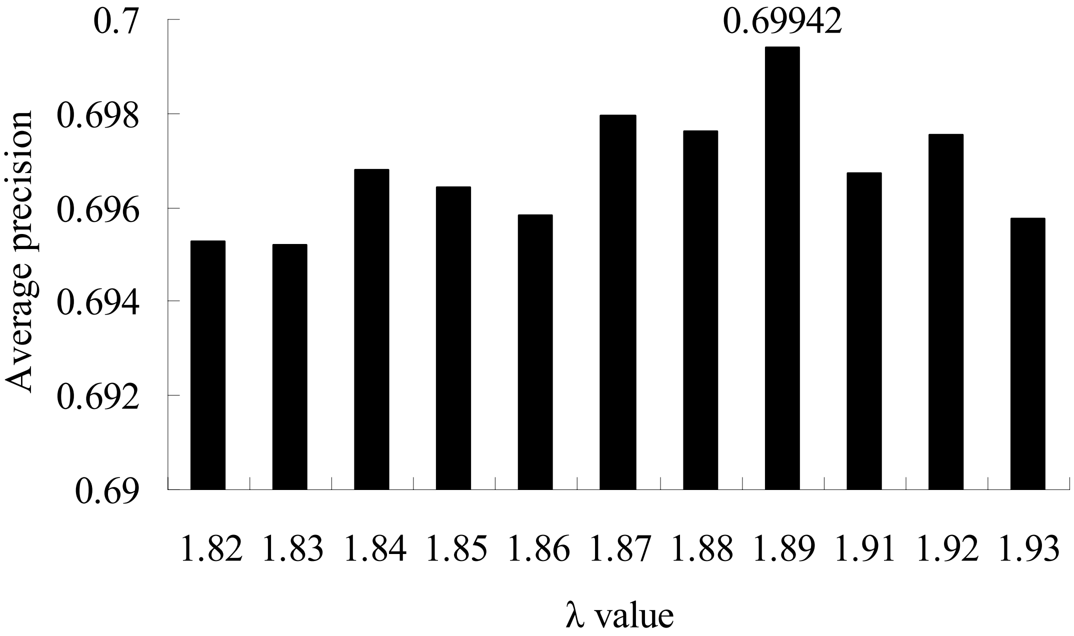
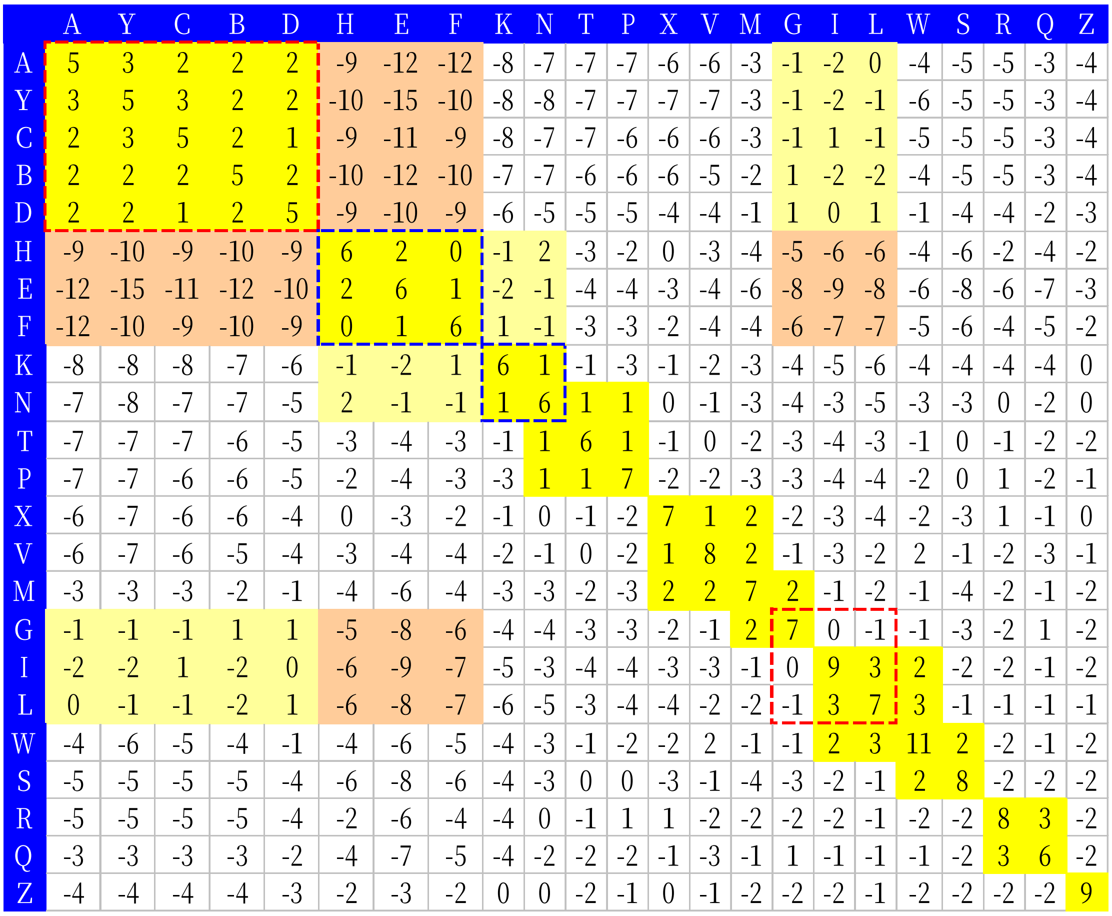


Figure 10. The average precision plot of PiSA-BLAST on 108 queries searching on SCOP 95 using various values of λ. The best performance λ value is 1.89, open gap penalty is 8, and extend gap penalty is 2.



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

Figure 11. The substitution matrix of 23 new codes. The scores on the diagonal cells are much higher than the scores on the non-diagonal cells. Red dot-square part (A, Y, C, B, and D) is the scores of aligning helix codes to helix codes and blue dot-square part (H, E, and F) is the scores of aligning strand codes to strand codes. The scores of aligning helix codes to strand codes are the smallest.

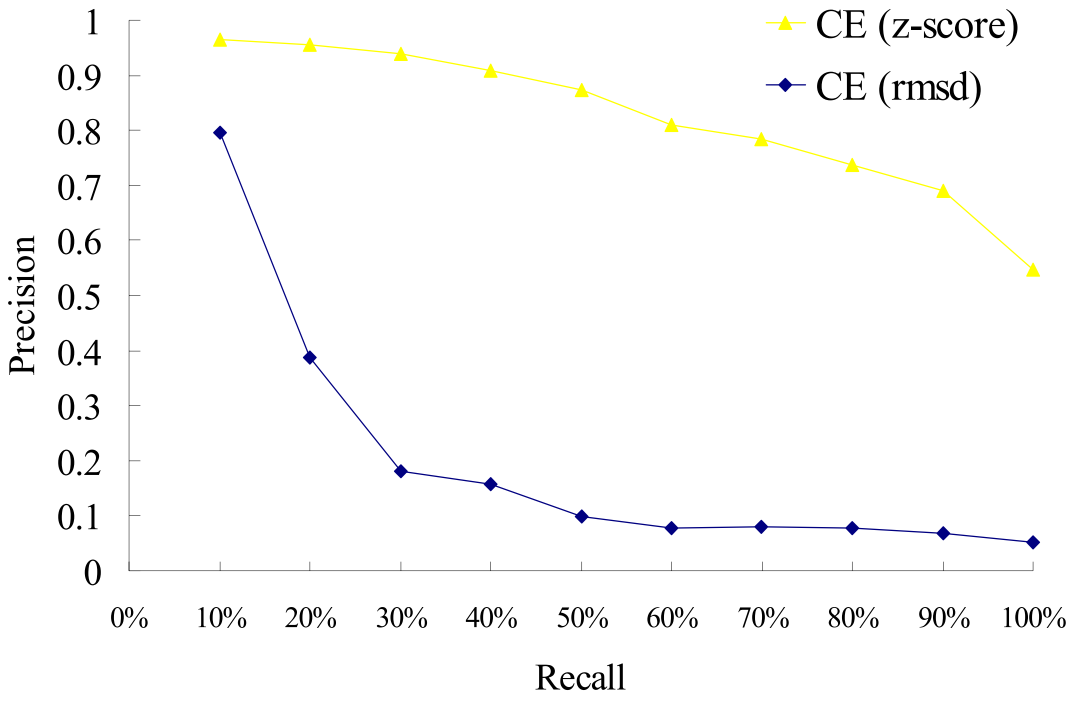


Figure 12. Recall-precision curves of CE using z-score and rmsd to order searching results on 108 queries searching the SCOP 95 database. The results of CE searching which are sorted by z-score are much more accurate than by rmsd.

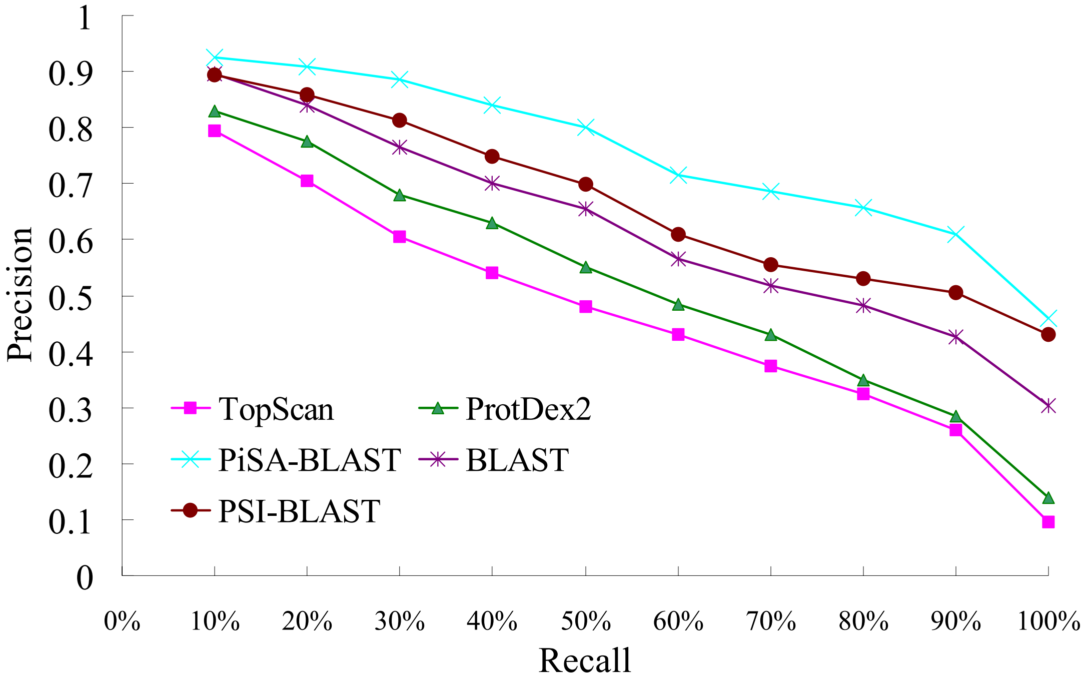


Figure 13. Recall-precision curves of five alignment tools for 108 queries on the large database of 33311 proteins indicated in Table 2. The results of ProtDex2 and TopScan, two fast structure alignment tools, are summarized from [14]. PiSA-BLAST is the best and TopScan is the worse among these five approaches.

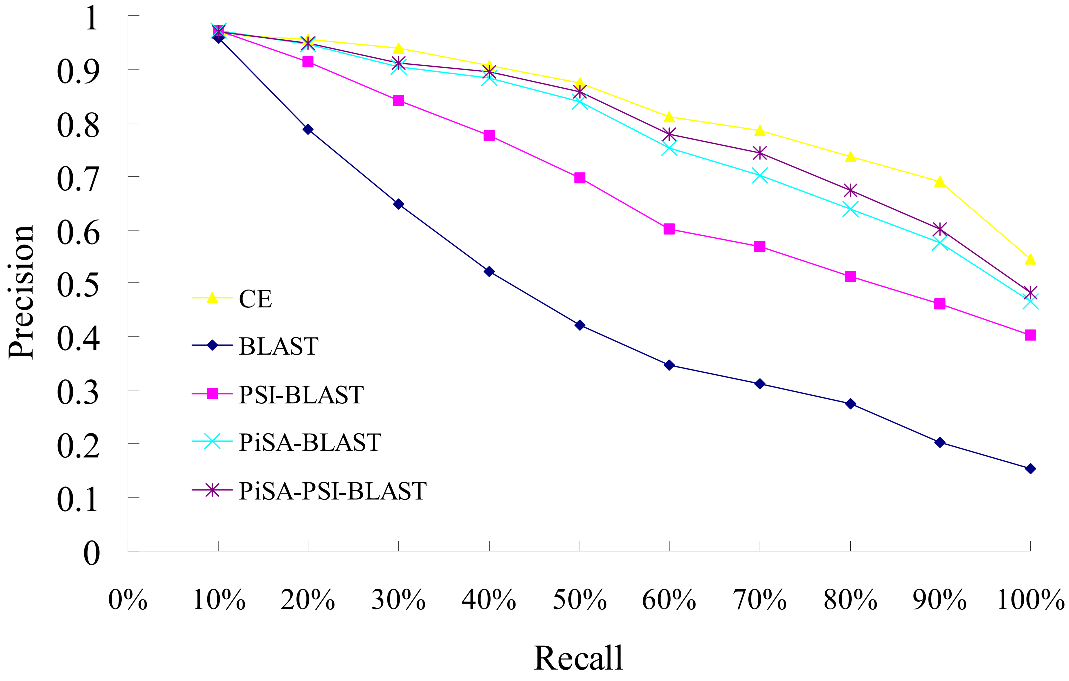


Figure 14. Recall-precision curves for 108 queries with CE, BLAST, PSI-BLAST, PiSA-BLAST and PiSA-PSI-BLAST on SCOP 95 database (ver 1.65). Accuracy of PiSA-BLAST closes the results of CE and PiSA-BLAST is about 34000 times fast than CE. PiSA-PSI-BLAST surprisingly only slightly improves PiSA-BLAST. In contrast, the performance of PSI-BLAST is much better than BLAST.

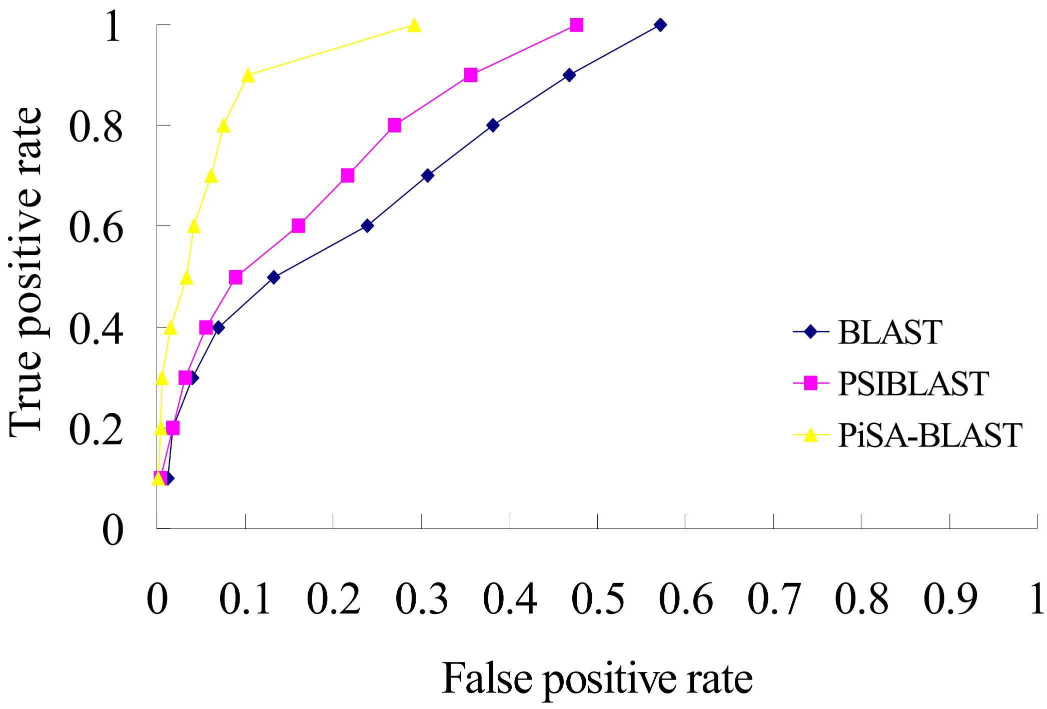


Figure 15. ROC curves of three tools performing 108 queries on the large database of 33311 proteins shown in Table 2. PiSA-BLAST can appear the accuracy more than other methods.

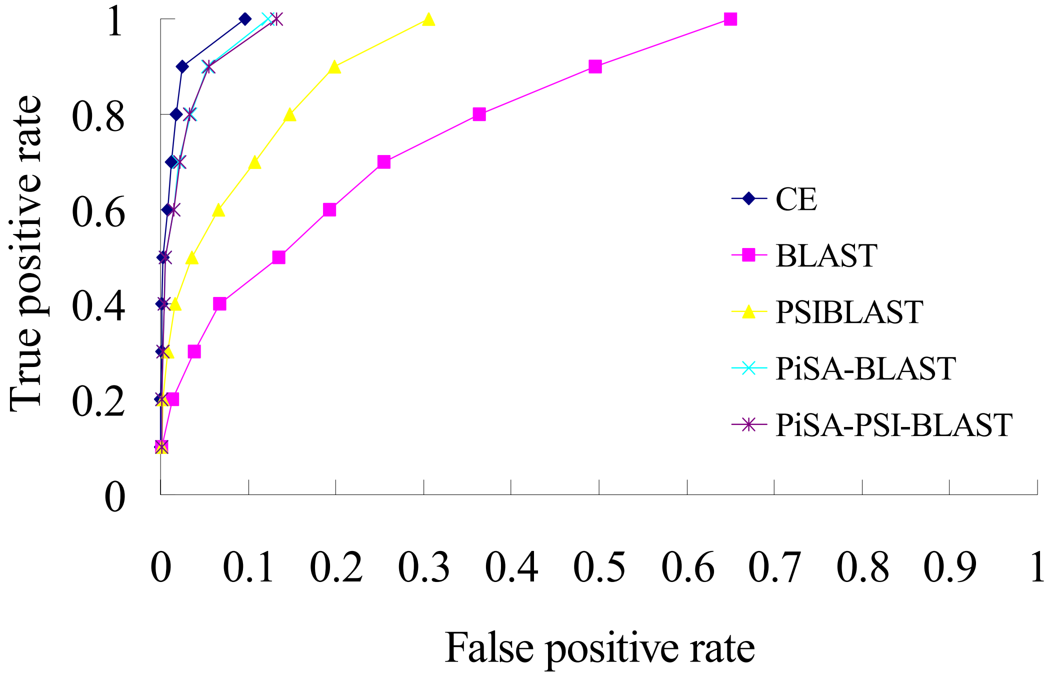
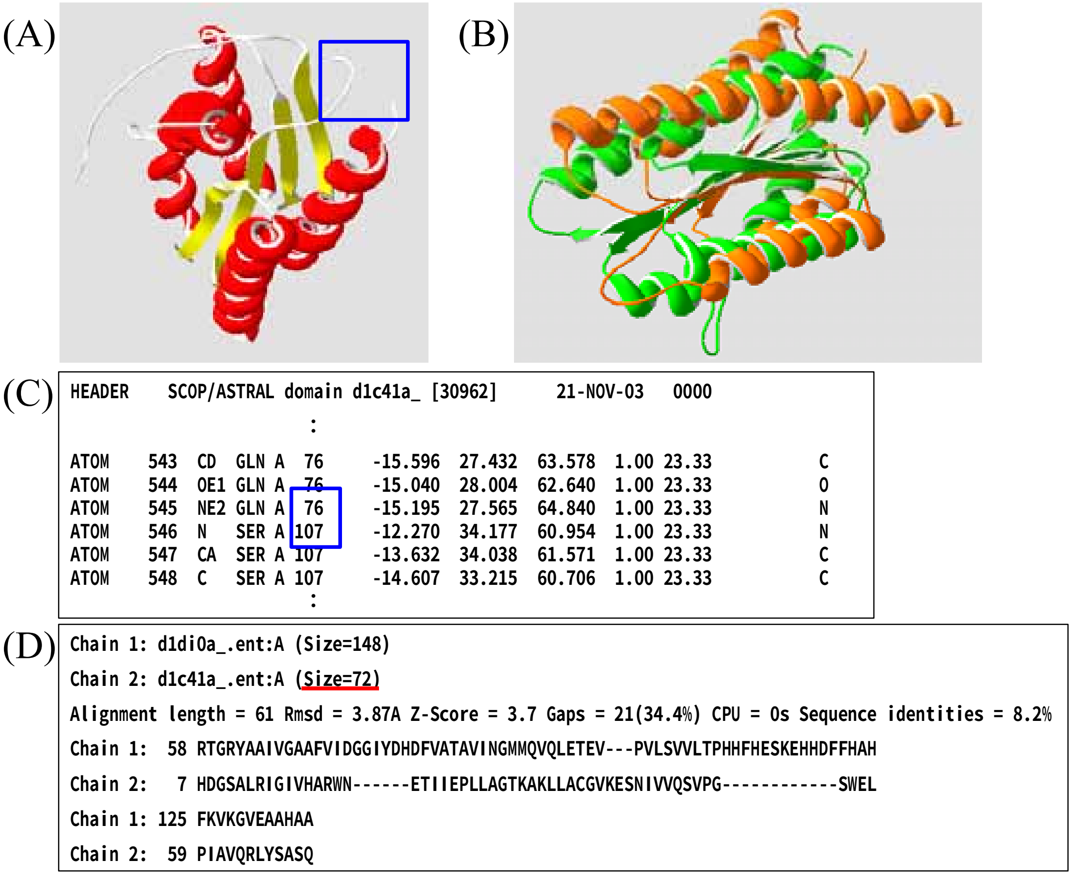


Figure 16. ROC curves of five tools perform 108 queries on SCOP 95% database. PiSA-BLAST and PiSA-PSI-BLAST can appear the performance close to CE and are more accurate than sequence alignment tools, BLAST and PSIBLAST.

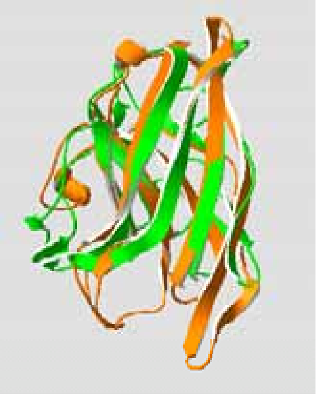


```
HEADER    SCOP/ASTRAL domain d1c41a_ [30962]      21-NOV-03   0000
                          :
ATOM     543  CD  GLN A  76     -15.596  27.432  63.578  1.00 23.33           C
ATOM     544  OE1 GLN A  76     -15.040  28.004  62.640  1.00 23.33           O
ATOM     545  NE2 GLN A  76     -15.195  27.565  64.840  1.00 23.33           N
ATOM     546  N   SER A 107     -12.270  34.177  60.954  1.00 23.33           N
ATOM     547  CA  SER A 107     -13.632  34.038  61.571  1.00 23.33           C
ATOM     548  C   SER A 107     -14.607  33.215  60.706  1.00 23.33           C
                          :
```

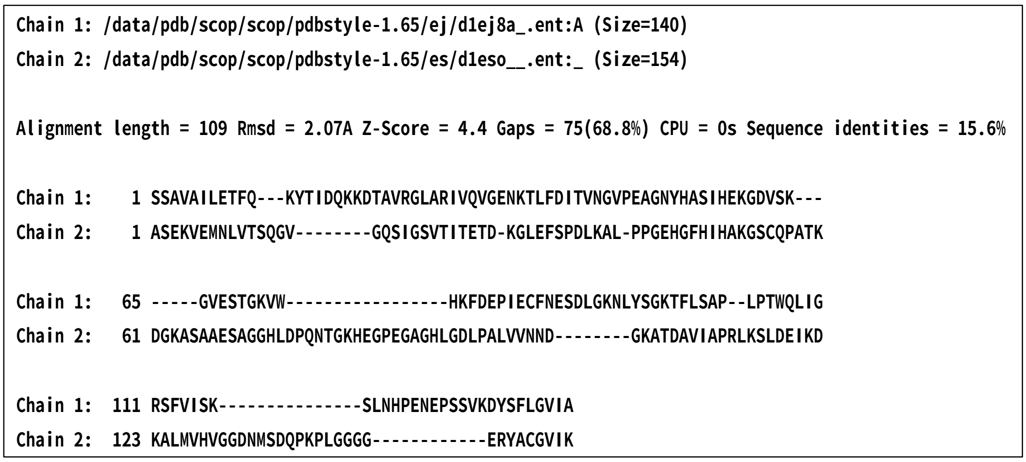
```
Chain 1: d1di0a_.ent:A (Size=148)
Chain 2: d1c41a_.ent:A (Size=72)
Alignment length = 61 Rmsd = 3.87A Z-Score = 3.7 Gaps = 21(34.4%) CPU = 0s Sequence identities = 8.2%
Chain 1:  58 RTGRYAAIVGAAFVIDGGIYDHDFVATAVINGMMQVQLETEV---PVLSVVLTPHHFHESKEHHDFFHAH
Chain 2:   7 HDGSALRIGIVHARWN------ETIIEPLLAGTKAKLLACGVKESNIVVQSVPG------------SWEL
Chain 1: 125 FKVKGVEAAHAA
Chain 2:  59 PIAVQRLYSASQ
```

Figure 17. The illustration of "chain-break" problem in CE alignment. (A) The 3D structure of subject protein "d1c41a_"; (B) the conformation of structure comparison of query "d1di0a_" and subject protein "d1c41a_" using CE; (C) the coordinate file of 3D structure in "d1c41a_"; (D) the alignment file of CE result. There is the condition of chain-break in subject protein "d1c41a_" shown with blue square in (A) and (C). The residue number is non-continuous from 76 to 107. The conformation of structure alignment of two proteins is slightly unsatisfied. Furthermore, the alignment length is shorter than the length of query protein and both Z-score and Rmsd is quite low as the alignment result in (D). Besides, we observed that CE determines the wrong length of the domain protein "d1c41a_". As shown in the red underline, the original length of "d1c41a_" is 165 but the size detected by CE is only 72 because of chain-break problem.

(A)



(B)



```
Chain 1: /data/pdb/scop/scop/pdbstyle-1.65/ej/d1ej8a_.ent:A (Size=140)
Chain 2: /data/pdb/scop/scop/pdbstyle-1.65/es/d1eso__.ent:_ (Size=154)

Alignment length = 109 Rmsd = 2.07A Z-Score = 4.4 Gaps = 75(68.8%) CPU = 0s Sequence identities = 15.6%

Chain 1:    1 SSAVAILETFQ---KYTIDQKKDTAVRGLARIVQVGENKTLFDITVNGVPEAGNYHASIHEKGDVSK---
Chain 2:    1 ASEKVEMNLVTSQGV--------GQSIGSVTITETD-KGLEFSPDLKAL-PPGEHGFHIHAKGSCQPATK

Chain 1:   65 -----GVESTGKVW-----------------HKFDEPIECFNESDLGKNLYSGKTFLSAP--LPTWQLIG
Chain 2:   61 DGKASAAESAGGHLDPQNTGKHEGPEGAGHLGDLPALVVNND--------GKATDAVIAPRLKSLDEIKD

Chain 1:  111 RSFVISK---------------SLNHPENEPSSVKDYSFLGVIA
Chain 2:  123 KALMVHVGGDNMSDQPKPLGGGG------------ERYACGVIK
```

Figure 18. The illustration of the problem of ordering the searching results by Z-score in CE alignment. (A) The conformation of structure comparison of query “#32 d1ej8a_” and subject protein “d1eso__” using CE; (B) the alignment file of CE result. The structures of query and subject proteins are similar and the rmsd is 2.07, but the Z-score is only 4.4. Therefore, the rank of the subject protein is 50 and behind 40 false positive proteins.

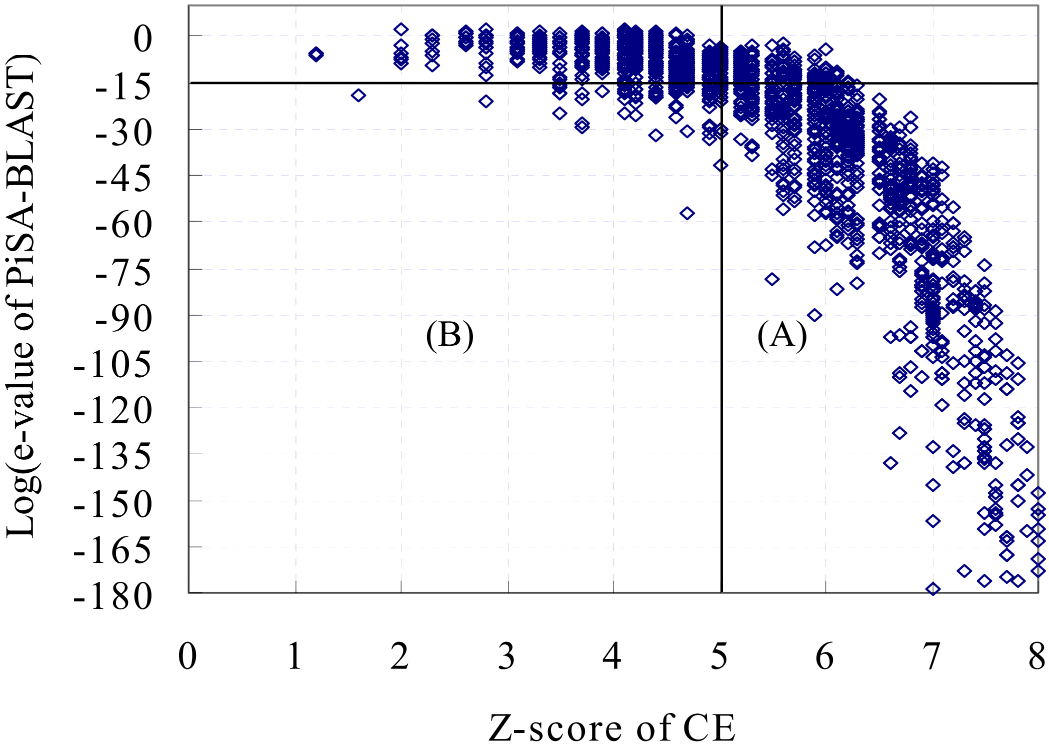


Figure 19. The relationship between e-value and structure similarity in PiSA-BLAST. The 1681 points in total on the plot mean every query and subject protein pairs searching in SCOP 95 database. There are 943 points in area (A) and only 79 points in area (B). PiSA-BLAST achieves 98.6% and 92.2% proteins whose Z scores are more than 4.0 and 5.0 when the e-value is less than 10^{-15}. PiSA-BLAST provides a significance estimate like e-value in BLAST to indicate what the performance is better.

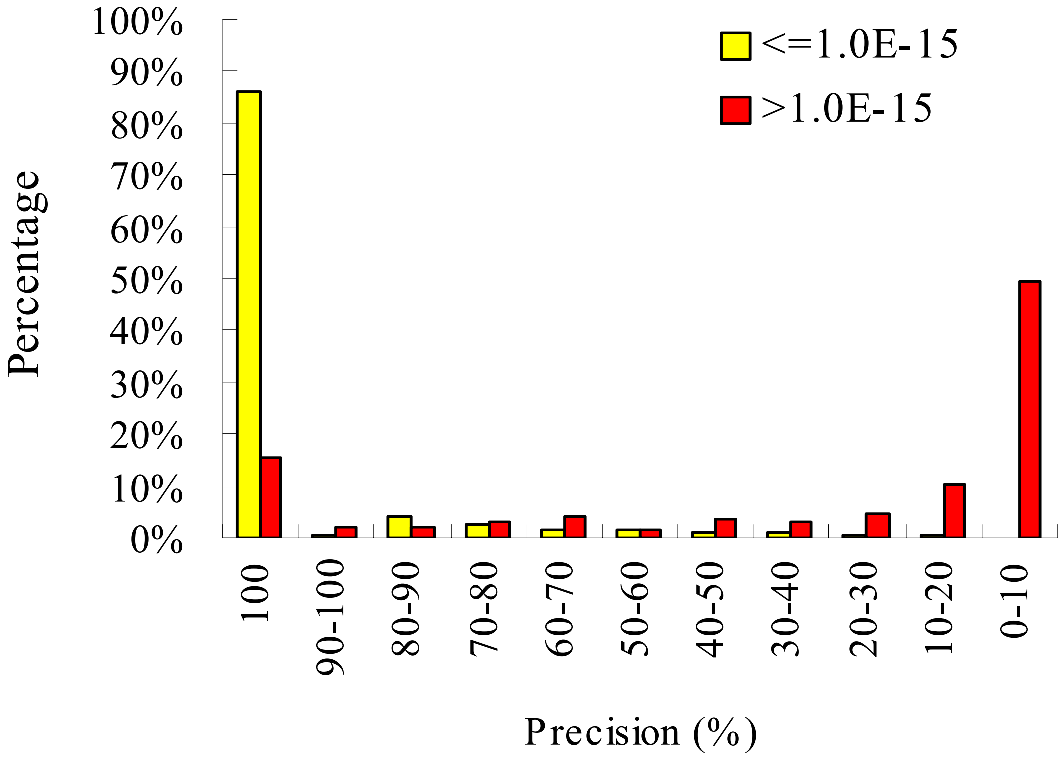


Figure 20. The relationship between e-value and precision in PiSA-BLAST. PiSA-BLAST performs 108 queries on the SCOP 95 database. The yellow bars mean that the distribution of e-value of PiSA-BLAST is less than 10^{-15} and red ones mean that the distribution of e-value is more than 10^{-15}. The protein pairs of precision with 80% and upper occupy 91% protein pairs at below 10^{-15} of e-value of PiSA-BLAST.

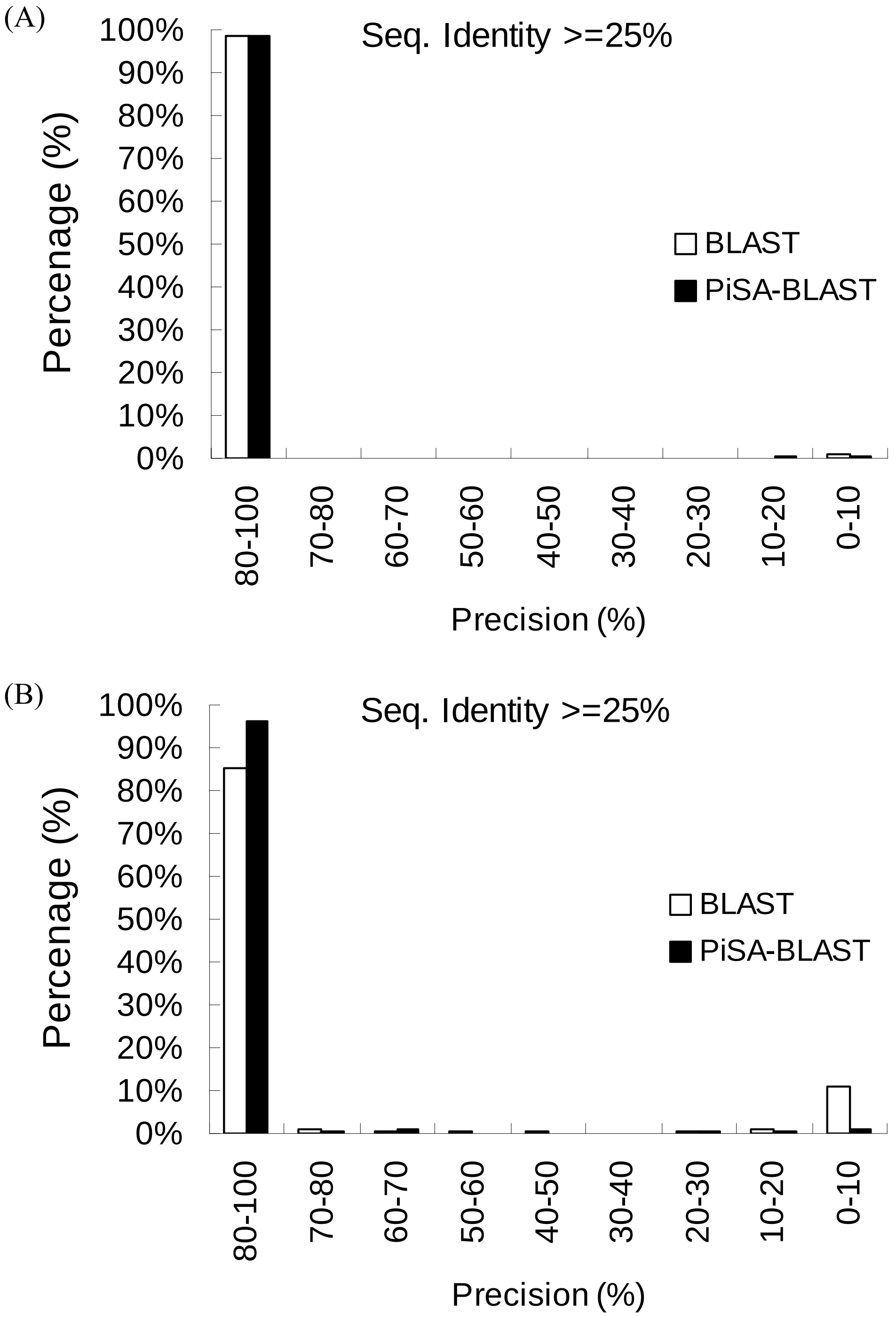


Figure 21. Comparison PiSA-BLAST with BLAST with high sequence identity (> 25%) on two databases: (A) the database with 33311 proteins shown in Table 2 and (B) the SCOP 95. PiSA-BLAST and BLAST have the similar performance.

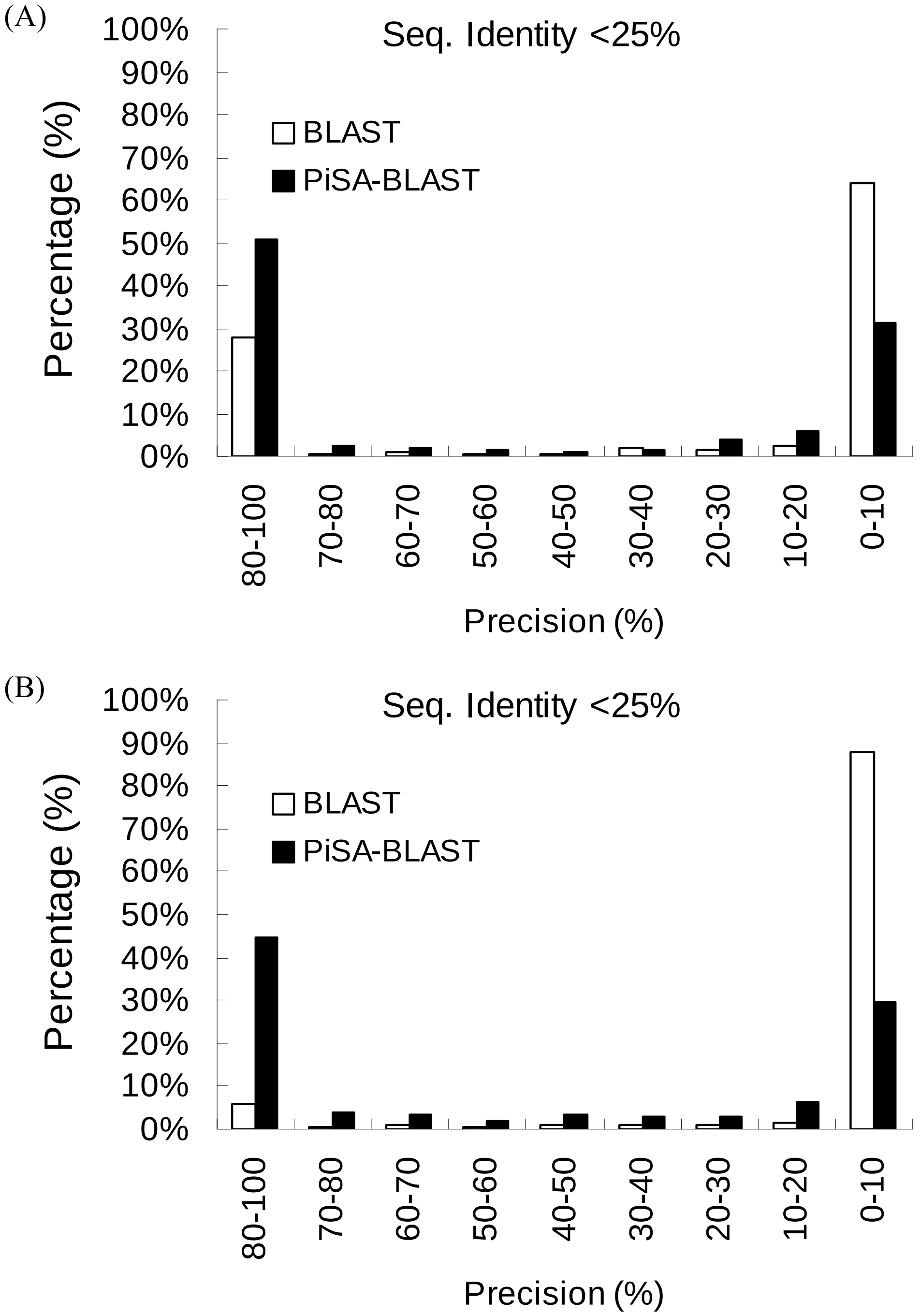


Figure 22. Comparison PiSA-BLAST with BLAST with low sequence identity (< 25%) on two databases: (A) the database with 33311 proteins shown in Table 2 and (B) the SCOP 95. PiSA-BLAST is much better than BALST for low sequence identity. The performance of BALST is more sensitive to the sequence identity than PiSA-BLAST do.

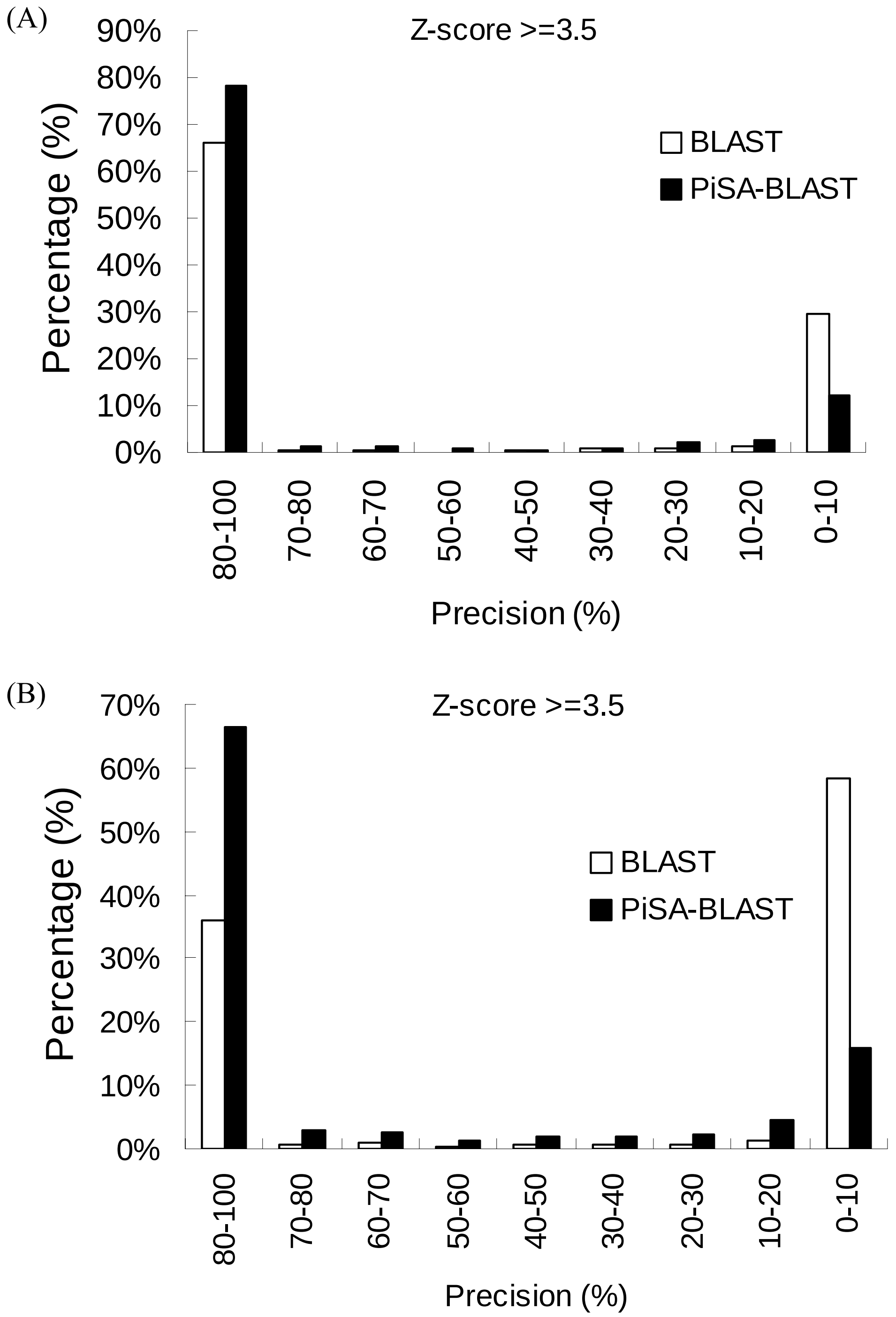


Figure 23. Comparison PiSA-BLAST with BLAST with high Z-score (> 3.5 by CE) on two databases: (A) the database with 33311 proteins shown in Table 2 and (B) the SCOP 95. PiSA-BLAST outperforms BLAST, especially, when the sequence identity is low.

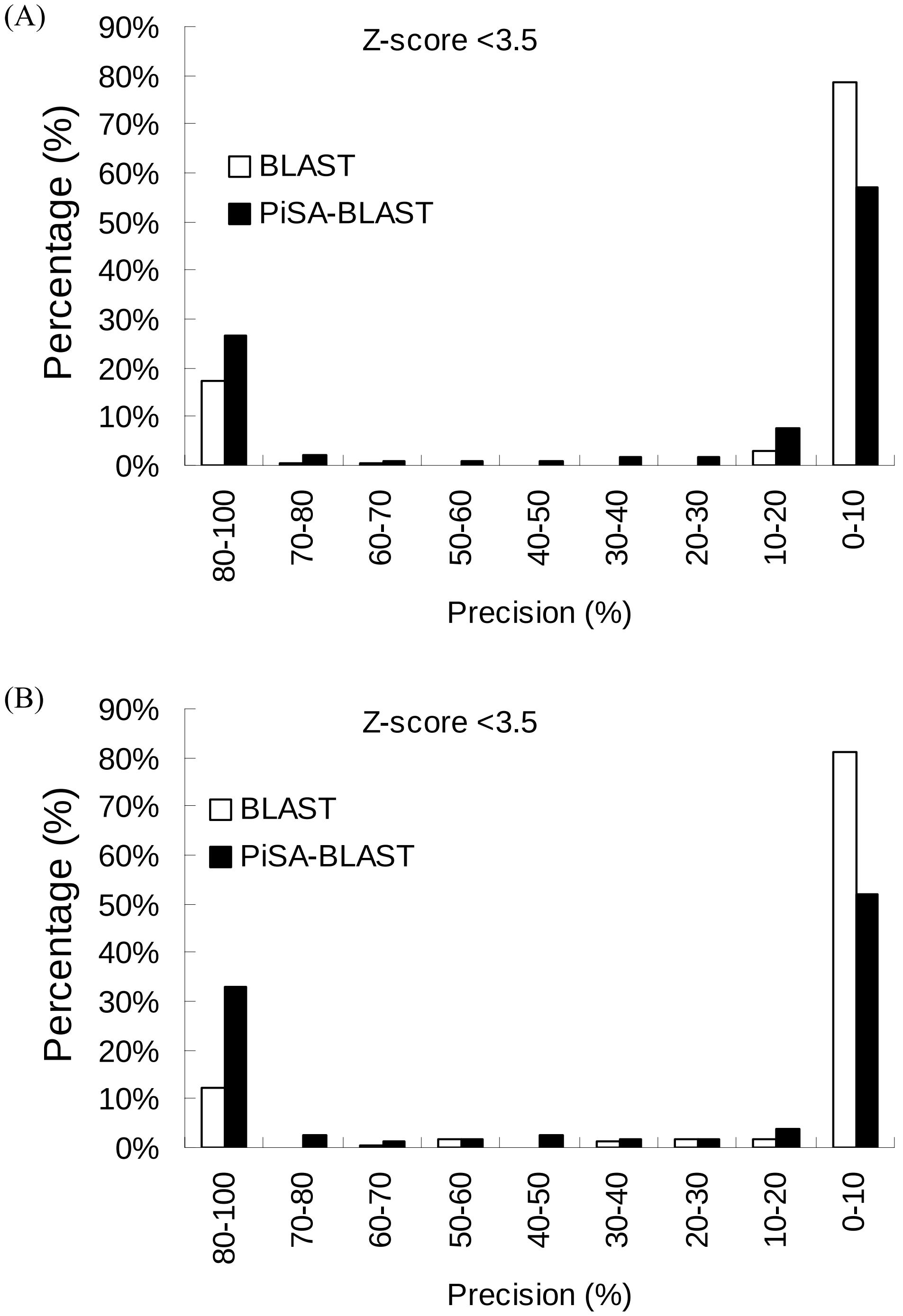
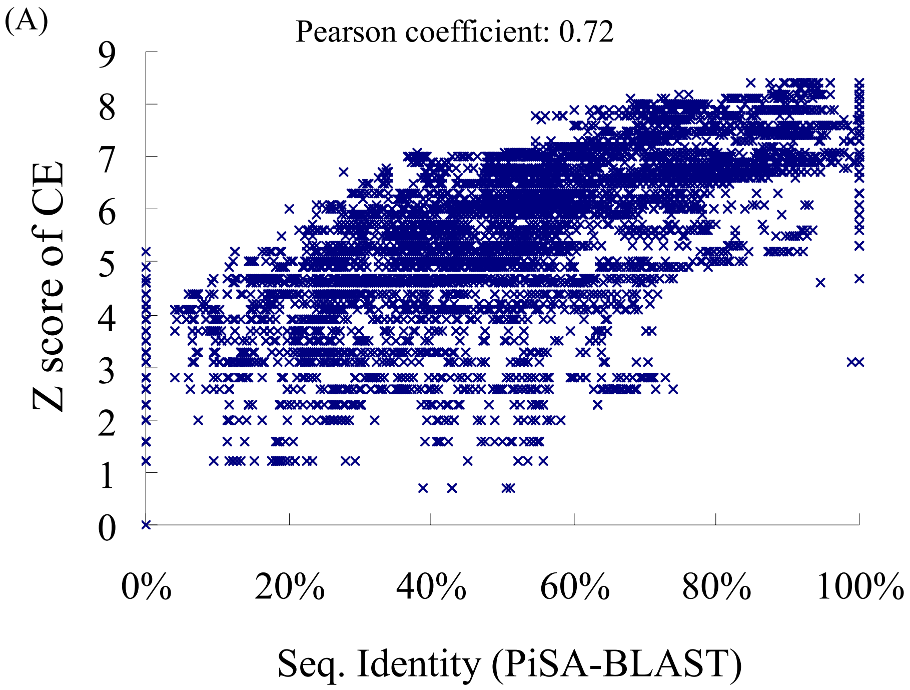


Figure 24. Comparison PiSA-BLAST with BLAST with low Z-score (< 3.5 by CE) on two databases: (A) the database with 33311 proteins shown in Table 2 and (B) the SCOP 95. PiSA-BLAST outperforms BLAST, especially, when the sequence identity is low. .



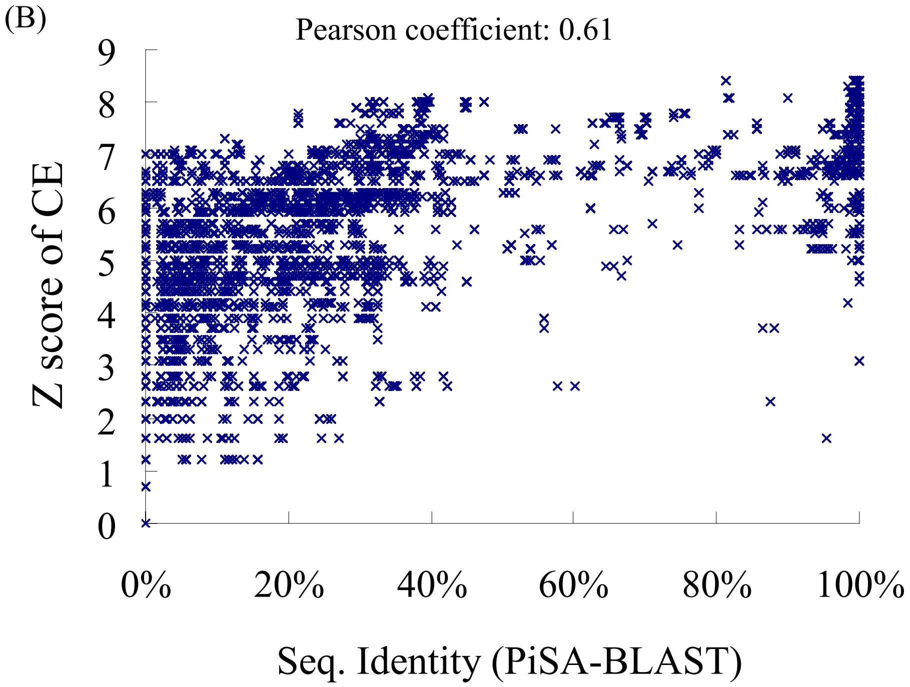


Figure 25. The correlations between Z-score (CE) and sequence identity calculated by (A) PiSA-BLAST and (B) BLAST. The correlation coefficient is 0.72 between encoded sequence identity of PiSA-BLAST and Z-score of CE, on the other hand, the correlation coefficient is 0.61 between amino acid sequence identity and Z-score.

(A)

```
 d1qe0a1 95 aa vs. d1nj1a1 127 aa
15.5% identity;
----IEENL---DLFIVTM-----GDQADRYAVKLLNHLRHNGIKADKDYLQRKIKGQMK--QADRLGAKFTIVIGDQELENNKIDVKNMTTGESETIELDALVEYFKK
    .  ..   .. :: .     .... .   .: ..:.  :...  :  .: :..  :  . .  :. . . :: ..::..   ..   :::. : .:... : ...
SGLCLPPDVAAHQVVIVPIIFKKAAEEVMEACRELRSRLEAAGFRVHLD--DRDIRAGRKYYEWEMRGVPLRVEIGPRDLEKGAAVISRRDTGEKVTADLQGIEETLRE
```

(B)

```
 Score =  108 bits (272), Expect = 1e-25, Identities = 40/93 (43%), Positives = 74/93 (79%), Gaps = 5/93 (5%)
Query: 4   MPFEEF----VTIDCBBGGDCDB-CACBBSRTNHFKHMSRKKIAACCDBACDSNEMPEEEHNIADDCSXNKHEFFGLSXNEFHKKCGQMLYDC 91
            +PFEEF    V  +++++ +C+B CA++BSRT HF H+ + + +++++++++S  MP+EEHN ++++S +KHEFFGLS++++HKK  QM++DC
Sbjct: 11  VPFEEFHHTQVQMAYYYBAYCYBACABDBSRTRHFEHXTQTNDDCBYYACDCSRKMPFEEHNADCGASQPKHEFFGLSRHHKHKKGWQMDCDC 103
```

(C)

```
Chain 1: d1qe0a1.ent:A (Size=95)
Chain 2: d1nj1a1.ent:A (Size=127)
Alignment length = 93 Rmsd = 1.74A Z-Score = 5.5 Gaps = 5(5.4%) CPU = 0s Sequence identities = 18.3%
Chain 1:    2 EENLDLFIVTMG-----DQADRYAVKLLNHLRHNGIKADKDYLQRKIKGQMKQADRLGAKFTIVIGDQELENNKIDVKNMTTGESETIELDALVE
Chain 2:    9 VAAHQVVIVPIIFKKAAEEVMEACRELRSRLEAAGFRVHLDDRDIRAGRKYYEWEMRGVPLRVEIGPRDLEKGAAVISRRDTGEKVTADLQGIEE
```

(D)

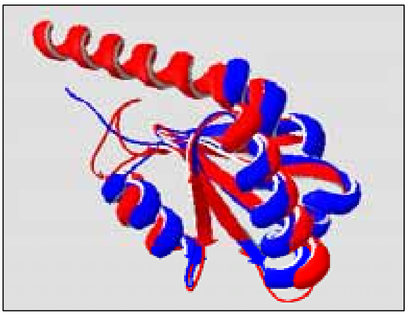
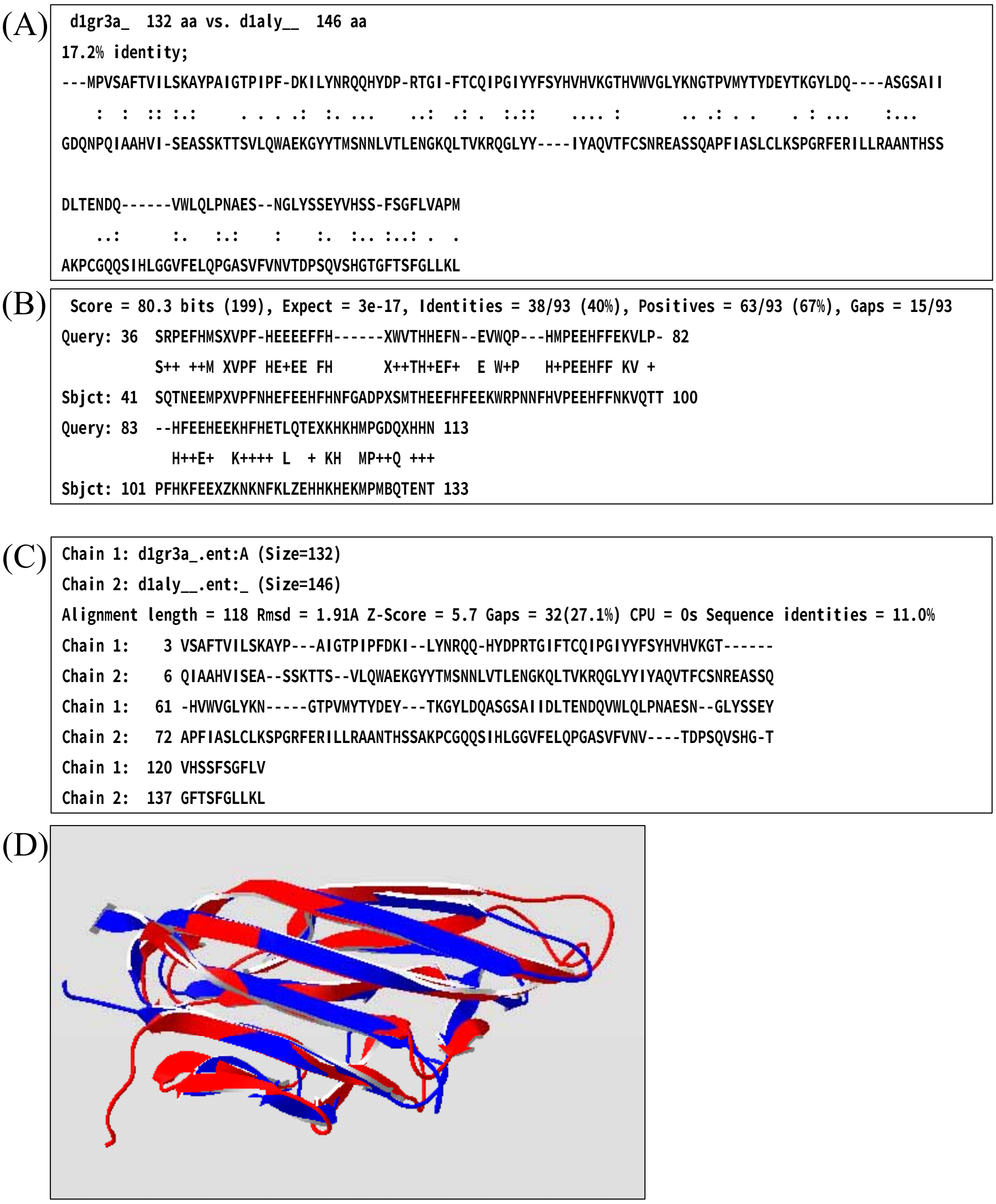


Figure 26. The results of FASTA, PiSA-BLAST and CE alignment to related domains: query protein “d1qe0a1” and subject protein “d1nj1a1”. (A) The sequence alignments in original amino acid by FASTA, (B) database searching with structural-encoded sequences by PiSA-BLAST, (C) structural alignment by CE and (D) the conformation of d1qe0a1 (blue) and d1nj1a1 (red) by CE. The sequence identity is 15.5% and the e-value of PiSA-BLAST is 10^{-25}. The Z-score of CE result is 5.5 and the conformation between query protein and subject protein is similar.



(A)

```
 d1gr3a_  132 aa vs. d1aly__  146 aa
17.2% identity;
---MPVSAFTVILSKAYPAIGTPIPF-DKILYNRQQHYDP-RTGI-FTCQIPGIYYFSYHVHVKGTHVWVGLYKNGTPVMYTYDEYTKGYLDQ----ASGSAII
    :  :  :: :.:     . . . .:  :. ...    ..:  .: .  :.::    .... :       .. .: . .    . : ...    :...
GDQNPQIAAHVI-SEASSKTTSVLQWAEKGYYTMSNNLVTLENGKQLTVKRQGLYY----IYAQVTFCSNREASSQAPFIASLCLKSPGRFERILLRAANTHSS

DLTENDQ------VWLQLPNAES--NGLYSSEYVHSS-FSGFLVAPM
    ..:      :.   :.:    :    :.  :.. :..: .  .
AKPCGQQSIHLGGVFELQPGASVFVNVTDPSQVSHGTGFTSFGLLKL
```

(B)

```
 Score = 80.3 bits (199), Expect = 3e-17, Identities = 38/93 (40%), Positives = 63/93 (67%), Gaps = 15/93
Query: 36  SRPEFHMSXVPF-HEEEEFFH------XWVTHHEFN--EVWQP---HMPEEHFFEKVLP- 82
           S++ ++M XVPF HE+EE FH      X++TH+EF+  E W+P   H+PEEHFF KV +
Sbjct: 41  SQTNEEMPXVPFNHEFEEHFHNFGADPXSMTHEEFHFEEKWRPNNFHVPEEHFFNKVQTT 100
Query: 83  --HFEEHEEKHFHETLQTEXKHKHMPGDQXHHN 113
             H++E+  K++++ L  + KH  MP++Q +++
Sbjct: 101 PFHKFEEXZKNKNFKLZEHHKHEKMPMBQTENT 133
```

(C)

```
Chain 1: d1gr3a_.ent:A (Size=132)
Chain 2: d1aly__.ent:_ (Size=146)
Alignment length = 118 Rmsd = 1.91A Z-Score = 5.7 Gaps = 32(27.1%) CPU = 0s Sequence identities = 11.0%
Chain 1:    3 VSAFTVILSKAYP---AIGTPIPFDKI--LYNRQQ-HYDPRTGIFTCQIPGIYYFSYHVHVKGT------
Chain 2:    6 QIAAHVISEA--SSKTTS--VLQWAEKGYYTMSNNLVTLENGKQLTVKRQGLYYIYAQVTFCSNREASSQ
Chain 1:   61 -HVWVGLYKN-----GTPVMYTYDEY---TKGYLDQASGSAIIDLTENDQVWLQLPNAESN--GLYSSEY
Chain 2:   72 APFIASLCLKSPGRFERILLRAANTHSSAKPCGQQSIHLGGVFELQPGASVFVNV----TDPSQVSHG-T
Chain 1:  120 VHSSFSGFLV
Chain 2:  137 GFTSFGLLKL
```

(D)

Figure 27. The results of FASTA, PiSA-BLAST and CE alignment to related domains: query protein “d1gr3a_” and subject protein “d1aly__”. (A) The sequence alignments in original amino acid by FASTA, (B) database searching with structural-encoded sequences by PiSA-BLAST, (C) structural alignment by CE and (D) the conformation of d1gr3a_(blue) and d1aly_(red) by CE. The sequence identity is 17.2% and the e-value of PiSA-BLAST is $3*10^{-17}$. The Z-score of CE result is 5.7 and the conformation between query protein and subject protein is similar.

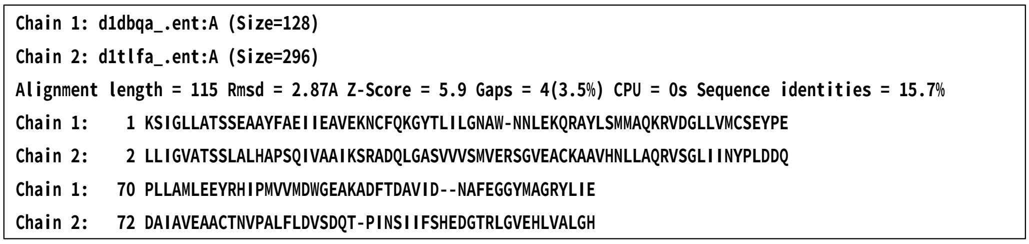
(A)

```
 d1dbqa_  276 aa vs. d1tlfa_  296 aa
25.8% identity;
KSIGLLATSSEAAYF-AEIIEAVEKNCFQKGYTLILGNAW-NNLEKQRAYLSMMAQKRVDGLLVMCSEYP---EPLLAMLEEYRHIPMVVMDWGEA
. .  .:::: : .  ..:. :...   : : ..... .  ...:  .: .  .  .::.::..   .::   .  .:.     ..: . .: ..
SLLIGVATSSLALHAPSQIVAAIKSRADQLGASVVVSMVERSGVEACKAAVHNLLAQRVSGLII---NYPLDDQDAIAVEAACTNVPALFLDVSDQ

KADFTDAVIDNAFEGGYMAGRYLIERGHREIGVIPGPAG------RLAGFMKAMEEAMIK-VPESWIVQGDFEPESGYRAMQQILSQPHRPTAVFC
     ...: .  .:  .. ..:.  ::..:... :: .      ::::. : . . .:. . :    .::.   ::..  .:.:..   :::..
TP--INSIIFSHEDGTRLGVEHLVALGHQQIALLAGPLSSVSARLRLAGWHKYLTRNQIQPIAER---EGDWSAMSGFQQTMQMLNEGIVPTAMLV

GGDIMAMGALCAADEMGLRVPQDVSLIGYDNVRNARYFTPALTTIHQPKDSLGETAFNMLLDRIVNKREEPQSIEVHP-RLIER-----------
..: ::.::. :  : ::::  :.:..:::.....  . : ::::.:    ::.:. . ::.  ... .  .. .. :  :..:
ANDQMALGAMRAITESGLRVGADISVVGYDDTEDSSCYIPPLTTIKQDFRLLGQTSVDRLLQ--LSQGQAVKGNQLLPVSLVKRKTTLAPNTQTAS

-RSVADGPF---RDYRR-
 :..::. .   :.  :
PRALADSLMQLARQVSRL
```

(B)

```
 Score =  253 bits (655), Expect = 5e-69
 Identities = 101/272 (37%), Positives = 211/272 (77%), Gaps = 20/272 (7%)

Query: 1   FEEHTVPGQTDLGGCBBBYYCDCDDDSRTFFHHEHETQTGYACDDCDCBYCD-SNFVPEE 59
           FEEH V++Q ++  ++B+++++++D+SRTF+ ++++ Q+  A+++++CB++D S+ VP+E
Sbjct: 2   FEEHFVTQQSGDYCBYBDBBBAYBDASRTFEKEFEFKQPXLAACBBYCBABDGSPNVPFE 61

Query: 60  HNTMPEKIDCBDBDGBRXGSN-NEEEEMWTPKQSQTKEEHKXMBDCYBBACBYYCBSRHM 118
           H T + +++C+++++B  GS+ + E++M T  Q Q   + K ++++++++CBY++BSRHM
Sbjct: 62  HXT-NKNLACACYYBB--GSTQTNEFHMQTKDQPQSRTHKKHGDCAAYCDCBYABBSRHM 118

Query: 119 PHFHHK----------DYDCBBDAACDSRNFFFIBGRFEFNQTBBDCAACYCACDQPDPK 168
           P+F+HK          +Y+++B++++DS++++     F++NQ+B++++A++C+++
Sbjct: 119 PEFEHKTTLQPDCACYCYCBYBLCDIDSQTKEV--PHFHKNQPBDBYCAABCYACBSQHE 176

Query: 169 FVPFEFMPBABGDCBDCBCBBSRTKVILZPHEEFMPNKLQMBQTLSTFXNKHFXMBBYYC 228
           FVPF+F++ +++DC+++B++BS+TKVILZP+EE+MPNKLQMBQTLS+F N +F M++++C
Sbjct: 177 FVPFHFVTIDDBDCYBDBDDBSQTKVILZPEEEKMPNKLQMBQTLSNFNNTEFHMACADC 236

Query: 229 DCA-BDBCDDLISHHXTHFEEEFFHKEFHTQM 259
           ++A ++++++    HX    EE+ +KEFHT +
Sbjct: 237 AAAYCAAYCGSRTRHXSK--EEKNEKEFHTLG 266
```

(C)



```
Chain 1: d1dbqa_.ent:A (Size=128)
Chain 2: d1tlfa_.ent:A (Size=296)
Alignment length = 115 Rmsd = 2.87A Z-Score = 5.9 Gaps = 4(3.5%) CPU = 0s Sequence identities = 15.7%
Chain 1:    1 KSIGLLATSSEAAYFAEIIEAVEKNCFQKGYTLILGNAW-NNLEKQRAYLSMMAQKRVDGLLVMCSEYPE
Chain 2:    2 LLIGVATSSLALHAPSQIVAAIKSRADQLGASVVVSMVERSGVEACKAAVHNLLAQRVSGLIINYPLDDQ
Chain 1:   70 PLLAMLEEYRHIPMVVMDWGEAKADFTDAVID--NAFEGGYMAGRYLIE
Chain 2:   72 DAIAVEAACTNVPALFLDVSDQT-PINSIIFSHEDGTRLGVEHLVALGH
```

(D)

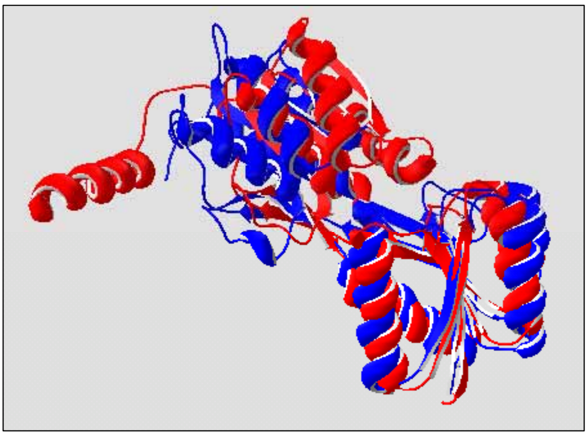


Figure 28. The results of FASTA, PiSA-BLAST and CE align with related domains: query protein “d1dbqa_” and subject protein “d1tlfa_”. (A) The sequence alignments in original amino acid by FASTA, (B) database searching with encoded sequence by PiSA-BLAST, (C) structural alignment and (D) the conformation of d1dbqa_ (blue) and d1tlfa_ (red) by CE. The sequence identity is 25.8%. The e-value in alignment of PiSA-BLAST is $5*10^{-69}$. The Z-score of CE result is 5.9 and the conformation between query protein and subject protein is similar.

(A)

```
 d1cjwa_ 166 aa vs. d1cm0a_ 162 aa
15.9% identity;
HTLPANEFRCL------TPEDAAGVFEIEREAFISVSGNCPLNLDEVQHFLTLCPEL-SLGWFVEGRLVAFIIGSLWDEERLTQESLALHRPRGHSAHL
...   ::. .       :.    .. .  .  .  : . :    :    :.. :.  .:. . .::    .::..          . .   .: .  .
KVI---EFHVVGNSLNQKPNKKILMWLVGLQNVF--SHQLPRMPKEYITRLVFDPKHKTLALIKDGR----VIGGI---------CFRMFPSQGFTEIV

HALAVHRSFRQQGKGSVLLWRYLHHVGAQPAVRRAVLMCEDALVPFYQRFGFHPAG--PCAIVVGSLTFTE----MHCSL---
   ::  . . .: :. :.  .:..   .  .   . . ..  . .... ::      : .  :: .   :    : : :
FC-AVTSNEQVKGYGTHLM-NHLKEYHIKHDILNFLTYADEYAIGYFKKQGFSKEIKIPKTKYVGYIKDYEGATLMGCELNPR
```

(B)

```
 Score =  106 bits (268), Expect = 6e-25
 Identities = 49/154 (31%), Positives = 101/154 (65%), Gaps = 27/154 (17%)
Query: 5   EHNTHKLGQM---YYYBDBGLILIBDLSGPKVPKLYBYBDCALLRQGGP--EFHEVWQTH 59
           +HN  +  Q+   Y+++++     ++   P  ++ +B+BDC+  + +GP  +FHEVW+TH
Sbjct: 4   FHNKPFGWQXXTGYCBABACYBYYACAAAPMQTFBBBBBDCD--QPBGPFFHFHEVWRTH 61
Query: 60  XPHEHHKNEMPGTNKIGGMQPEEGQPXPHFEVPHNKKCYQXQSQMAAACAYCBACDCDQS 119
           +P +H+K              E++   P F++P NKK++Q+QSQMA+++++CB+CD++ S
Sbjct: 62  VPFHHNK--------------EHMASQPFFHXPXNKKYCQMQSQMACBACACBYCDDBYS 107
Query: 120 GQHXPHXKHKTIBQGBDDGGSQNKHKVTFQPETW 153
            +HXP+ +++ ++Q+B++++S++K     QPE +
Sbjct: 108 -RHXPENNNFKLAQMBBBBBSRHK-----QPEXS 135
```

(C)

```
Chain 1: d1cjwa_.ent:A (Size=166)
Chain 2: d1cm0a_.ent:A (Size=163)
Alignment length = 135 Rmsd = 2.42A Z-Score = 5.5 Gaps = 50(37.0%) CPU = 0s Sequence identities = 14.8%
Chain 1:    5 ANEFRCLTP-----------EDAAGVFEIEREAFISVSGNCPLNLDEVQHFLTLCPELSLGWFVEGRLVA
Chain 2:    2 VIEFHVVGNSLNQKPNKKILMWLVGLQNVFSHQL--PRMPKEYITRLVFDPKH----KTLALIKDGRVIG
Chain 1:   64 FIIGSLWDEERLTQESLALHRPRGHSAHLHALAVHRSFRQQGKGSVLLWRYLHHVGAQPAVRRAVLMCED
Chain 2:   66 GICFRMF--------------PSQGFTEIVFCAVTSNEQVKGYGTHLMNHLKEYHIKHD-ILNFLTYADE
Chain 1:  134 ALVPFYQRFGFHPAGPC------------AIVVGSLTFTEMHCSL
Chain 2:  121 YAIGYFKKQGFSKE--IKIPKTKYVGYIKDYE----GATLMGCEL
```

(D)

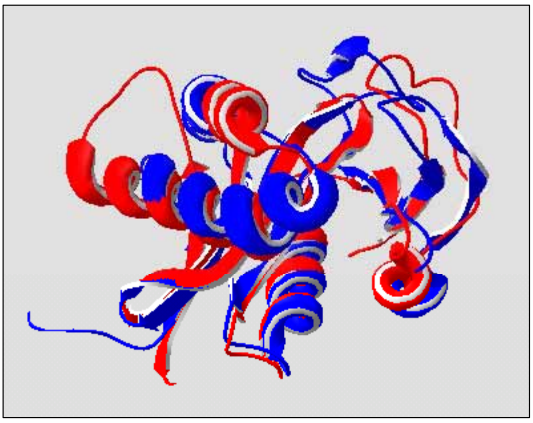


Figure 29. The results of FASTA, PiSA-BLAST and CE align with related domains: query protein “d1cjwa_” and subject one “d1cm0a_”. (A) The sequence alignments in original amino acid by FASTA, (B) database searching with encoded sequence by PiSA-BLAST, (C) structural alignment and (D) the conformation of d1cjwa_ (blue) and d1cm0a_ (red) by CE. The sequence identity between these two proteins is 15.9% and the e-value of PiSA-BLAST

is $6*10^{-25}$. The Z-score of CE result is 5.5 and the conformation between query protein and subject protein is similar.

(A)

```
 Score = 39.5 bits (92), Expect = 4e-05, Identities = 19/63 (30%), Positives = 40/63 (63%), Gaps = 15/63 (23%)
Query: 4   IDCDCBBYYAQPMS-KFKIEDCAA-------------DSRTNCDCBBDLBACACDSQMPF-FK 51
            ID+++BB+++QP S  FK+E++A+             D +    D++++++++++  QMPF FK
Sbjct: 9   IDYCBBBDABQPDSRHFKLEBDACBDALLPQMQMDGGDTQXMGDIGDBDGBBDIMPQMPFHFK 71
```

(B)

```
Chain 1: d1mkma1.ent:A (Size=75)
Chain 2: d1e17a_.ent:A (Size=90)
Alignment length = 59 Rmsd = 2.39A Z-Score = 3.5 Gaps = 26(44.1%) CPU = 0s Sequence identities = 10.2%
Chain 1:    4 LKKAFEILDFIVKNPG-DVSVSEIAEKFN----------------MSVSNAYKYMVVLEEKGFVLRKKD------KRYVPGYKLI
Chain 2:    9 QSYAELISQAIESAPEKRLTLAQIYEWMVRTVPYFKDKGDSNSSAGWKNSIRHNLSL---HSKFIKVHNEATGKSSWWMLNPEGG
```

(C)

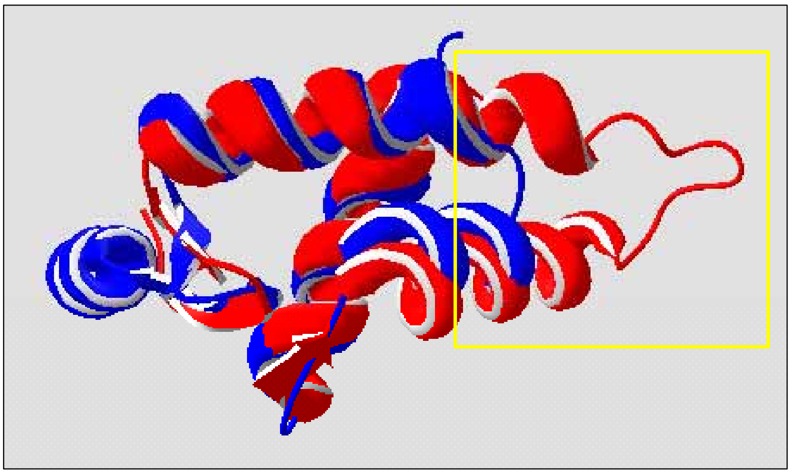
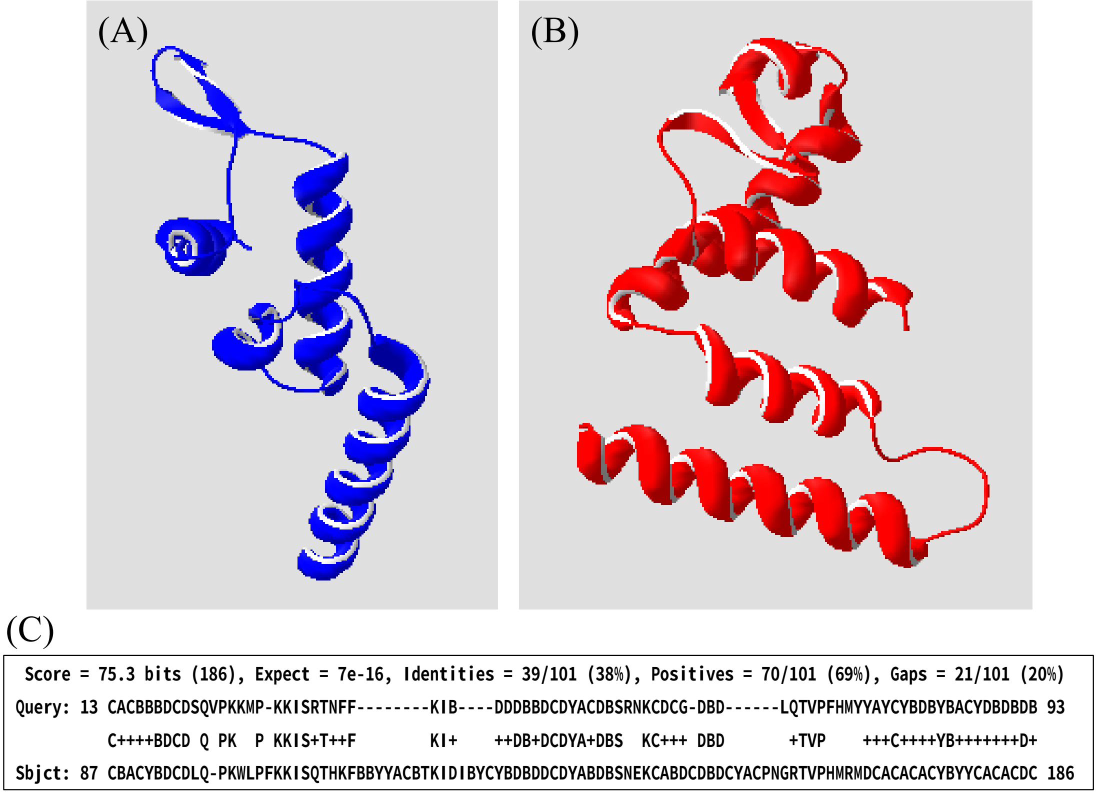


Figure 30. A bad case in our method for comparison of two related domain proteins, d1mkma1 and d1e17a_. The conformation between query protein “d1mkma1_” in blue and subject one “d1e17a_” in red is similar. However, there are the long structural gaps at yellow square in protein “d1e17a”. These gaps are necessary for structural comparison, but our alignment tool does not allow long gaps to exist. Because there are critical gap open and extension penalties in sequence alignment, the alignment score of PiSA-BLAST is low in this case.



```
 Score = 75.3 bits (186), Expect = 7e-16, Identities = 39/101 (38%), Positives = 70/101 (69%), Gaps = 21/101 (20%)
Query: 13 CACBBBDCDSQVPKKMP-KKISRTNFF--------KIB----DDDBBDCDYACDBSRNKCDCG-DBD------LQTVPFHMYYAYCYBDBYBACYDBDBDB 93
          C++++BDCD Q PK  P KKIS+T++F        KI+    ++DB+DCDYA+DBS  KC+++ DBD       +TVP    +++C++++YB+++++++D+
Sbjct: 87 CBACYBDCDLQ-PKWLPFKKISQTHKFBBYYACBTKIDIBYCYBDBDDCDYABDBSNEKCABDCDBDCYACPNGRTVPHMRMDCACACACYBYYCACACDC 186
```

Figure 31. A false positive example in PiSA-BLAST for comparison of two non-related domain proteins, (A) d1jbga_ and (B) d1pk5a_. The SCOP sccs id of "d1jbga_" is a.6.1.3 and SCOP id of "d1pk5a_" is a.123.1.1. All of these secondary structures are four to five helices and two short strands. The compositions of secondary structure are similar, but the three-dimensional conformation is quite different. The e-value in PiSA-BLAST alignment is 7*10^{-16} and the rank of the subject protein is 3.

(A)

```
 Score = 98.9 bits (248), Expect = 1e-21, Identities = 42/132 (31%), Positives = 89/132 (67%), Gaps = 23/132
(17%)
Query: 36  LYBYBDCALLRQGGPEFHEVWQTHXPHEHHKNEMPGTNKIGGMQPEEGQPXPHF--EVPHNKKCYQXQSQMAAACAYCBACDCDQSGQHXPHX
            +Y++++++ LR  ++EFH+VW +H+P++++            M++EEG  +  +  +VP++KKC+QXQSQM+++++YC+++D +     +P+
Sbjct: 39  DYYCCYYYALRMAQNEFHHVWLPHVPEFEEE-----------MRTEEGLSMVRKPKHVPNHKKCAQXQSQMBYYYBYCACADLG--LLPMPEH
            KHKT---IBQGBDDGGSQNKHKVTFQPETWRPEFKHEEF 162
            +HK     + ++++++S+NK+     +ETW P F +EEF
            FHKKAQHMDCBYYBBBSRNKN----XTETWLP-FHEEEF 152
```

(B)

```
Chain 1: pdb1cjw.ent:A (Size=166)
Chain 2: pdb1wwz.ent:B (Size=159)
Alignment length = 139 Rmsd = 2.35A Z-Score = 5.7 Gaps = 39(28.1%) CPU = 0s Sequence identities = 16.5%
Chain 1:    5 ANEFRCLTPE---DAAGVFEIEREAFISVSG-------NCPLNLDEVQHFLTLCPELSLGWFVEGRLVAFIIGS-LWDE--ERLTQESLA
Chain 2:    3 EIKIEKLKKLDKKALNELIDVYMSGYEGLEEYGGEGRDY---ARNYIKWCWKKASDGFFVAKVGDKIVGFIVCDKDWFSKYEG-------
Chain 1:   82 LHRPRGHSAHLHALAVHRSFRQQGKGSVLLWRYLHHVGAQPAVRRAVLMCEDA---LVPFYQRFGFHPAGPCAIVVGSLTFTEMHCSL
Chain 2:   83 -----RIVGAIHEFVVDKKFQGKGIGRKLLITCLDFLGK--YNDTIELWVGEKNYGAMNLYEKFGFKKVGKSG------IWVRMIKRQ
```

(C)

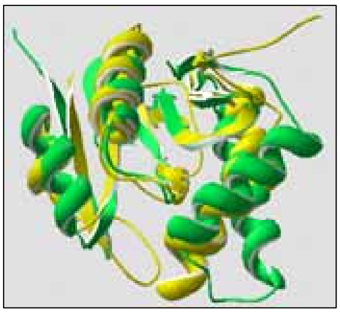


Figure 32. A practical application of fold assignment using PiSA-BLAST. The results of (A) encoded sequence alignment, (B) structural alignment and (C) 3-dimensional conformation between the query protein “1cjwA” and “1wwzB”. The protein “1wwz” is published on 01-Feb-05 and is not assigned in SCOP and CATH currently. PiSA-BLAST is used to assign the fold of the protein “1wwz”, highly similarly to protein “1cjw”, to SCOP sccs id: “d.108.1.1”. The e-value of PiSA-BLAST alignment between “1cjwA” and “1wwzB” is less than 10^{-15}. Then, we performed CE for detail structure alignment and the Z-score of CE alignment is 5.7. Therefore, we suggested that the protein “1wwzB” is assigned the fold of the protein “1cjwA”.

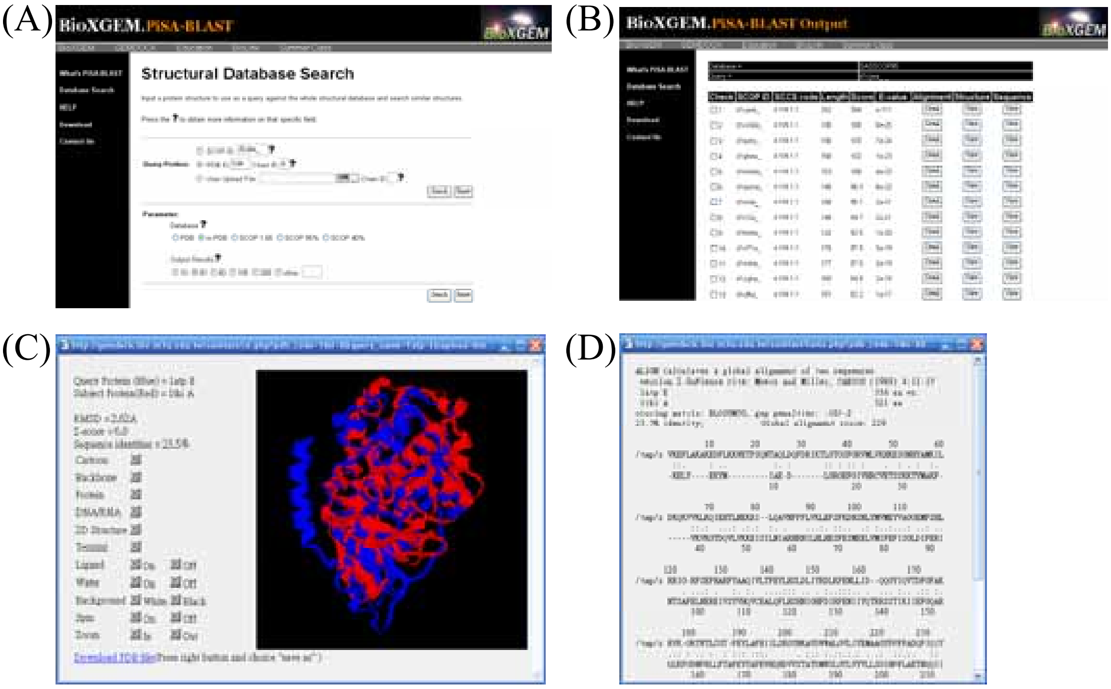


Figure 33. The illustration of PiSA-BLAST web service. (A) Query interface: there are three kinds of query formats: PDB code, SCOP code, and users' upload 3D structure. The searching databases includes PDB, nr-PDB, SCOP all, SCOP 95, and SCOP 40; (B) The query results includes protein ID, scores, and e-value; (C) Structure alignments between query and subject proteins using CE; (D) Amino acid sequence alignments between query and subject proteins using FASTA program. PiSA-BLAST is available on http://gemdock.life.nctu.edu.tw/pisa-blast/.

REFERENCES

1. Matthews, B.W. and M.G. Rossmann, *Comparison of protein structures.* Methods in Enzymology, 1985. **115**: p. 397-420.
2. Jain, A.K. and R.C. Dubes, *Algorithms for Clustering Data*. 1988, New Jersey: Prentice Hall: Englewood Cliffs.
3. 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.
4. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **215**: p. 403-410.
5. Chothia, C. and A.M. Lesk, *The relation between the divergence of sequence and structure in proteins.* The EMBO Journal, 1986. **5**: p. 823-826.
6. Lesk, A.M. and C. Chothia, *How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins.* Journal of Molecular Biology, 1980. **136**: p. 225-270.
7. Sander, C. and R. Schneider, *Database of homology-derived protein structures and the structural meaning of sequence alignment.* Proteins: Structure, Function, and Bioinformatics, 1991. **9**: p. 56-68.
8. Rost, B., *Twilight zone of protein sequence alignments.* Protein Engineering, 1999. **12**: p. 85-94.
9. Holm, L. and C. Sander, *Protein structure comparison by alignment of distance matrices.* Journal of Molecular Biology, 1993. **233**: p. 123-138.
10. Madej, T., J.F. Gibrat, and S.H. Bryant, *Threading a database of protein cores.* Proteins: Structure, Function, and Bioinformatics, 1995. **23**: p. 356-369.
11. Gibrat, J.F., T. Madej, and S.H. Bryant, *Surprising similarities in structure comparison.* Current Opinion in Structural Biology, 1996. **6**: p. 377-385.
12. 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.
13. Martin, A.C., *The ups and downs of protein topology; rapid comparison of protein structure.* Protein Engineering, 2000. **13**: p. 829-837.
14. Aung, Z. and K.L. Tan, *Rapid 3D protein structure database searching using information retrieval techniques.* Bioinformatics, 2004. **20**: p. 1045-1052.
15. Hubbard, T.J.P., et al., *SCOP: a structural classification of proteins database.* Nucleic Acids Research, 1997. **25**(1): p. 236-239.
16. 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.

17. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**: p. 2577-2637.

18. Paul, J.B. and N.D. M., *A method for registration of 3-D shapes.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992. **14**: p. 239-256.

19. 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.

20. Fano, R.M., *In Transmission of information; A Statistical Theory of Communications*. 1961, New York.: Wiley.

21. Pearson, W.R., *Rapid and sensitive sequence comparison with FASTP and FASTA.* Methods in Enzymology, 1990. **183**: p. 63-98.

22. 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.